DOCUMENT RESUME

ED 366 294 IR 016 488

TITLE	National Collaboratories: Applying Information Technology for Scientific Research.
INSTITUTION	National Academy of Sciences - National Research Council, Washington, DC. Computer Sciences and Telecommunications Board.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	ISBN-0-309-04848-6
PUB DATE	93
CONTRACT	CDA-9021110
NOTE	116p.; Prepared by the Committee on a National Collaboratory.
AVAILABLE FROM	National Academy Press, 2101 Constitution Ave., N.W., Washington, DC 20418.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Cooperation; *Information Technology; Molecular Biology; Oceanography; Participative Decision Making; *Scientific Research; Space Sciences
IDENTIFIERS	National Research Council; National Science Foundation

ABSTRACT

The Committee on a National Collaboratory: Establishing the User-Developer Partnership was charged to study and report on the need for and potential of information technology to support collaboration in the conduct of scientific research. To do this, the committee focused on three discrete areas of scientific investigation: (1) oceanography, in which the difficulty and expense of gathering data and the interdependence of modelers and experimentalists provide motivation for greater collaboration; (2) space physics, which has of necessity used extensive computational technology in the analysis of data collected by cooperatively fielded space- and ground-based instruments; and (3) gene mapping and sequencing research that has led to construction of and reliance on massive databases. The committee's investigations suggested technical requirements and social and practical issues that must be considered and dealt with as part of the process of initiating a national collaboratory program in support of scientific research. In conducting this study, the committee sought to identify common information technology needs that cross disciplines; identify specific information technology needs in three particular fields of science; increase awareness of the utility of information technology for the conduct of scientific research; and identify goals, objectives, and costs of developing collaboratories. (TMK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

ED 366 294

NATIONAL COLLABORATORIES

Applying Information Technology for Scientific Research

BEST COPY AVAILABLE

12016 488

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Lisa Magnino
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Computer Science and Telecommunications Board

NATIONAL COLLABORATORIES

Applying Information Technology for Scientific Research

Committee on a National Collaboratory:
Establishing the User-Developer Partnership
Computer Science and Telecommunications Board
Commission on Physical Sciences, Mathematics, and Applications
National Research Council

National Academy Press
Washington, D.C. 1993

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competencies and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

Support for this project was provided by the National Science Foundation (Grant No. CDA-9021110).

Library of Congress Catalog Card Number 93-83795
International Standard Book Number 0-309-04848-6

Copyright 1993 by the National Academy of Sciences. All rights reserved.

Available from:

National Academy Press
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

B-122

Printed in the United States of America

COMMITTEE ON A NATIONAL COLLABORATORY:
DEVELOPING THE USER-DEVELOPER PARTNERSHIP

VINTON G. CERF, Corporation for National Research Initiatives, *Chairman*
ALASTAIR G.W. CAMERON, Harvard College Observatory, *Vice-Chairman*
JOSHUA LEDERBERG, Rockefeller University
CHRISTOPHER T. RUSSELL, University of California at Los Angeles
BRUCE R. SCHATZ, University of Arizona
PETER M.B. SHAMES, California Institute of Technology
LEE S. SPROULL, Boston University
ROBERT A. WELLER, Woods Hole Oceanographic Institution
WILLIAM A. WULF, University of Virginia

Staff

MARJORY S. BLUMENTHAL, Director
MONICA KRUEGER, Staff Officer
ARTHUR L. McCORD, Project Assistant (through February 1993)
LESLIE M. WADE, Project Assistant

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

WILLIAM A. WULF, University of Virginia, *Chairman*
RUZENA BAJCSY, University of Pennsylvania
DAVID J. FARBER, University of Pennsylvania
SAMUEL H. FULLER, Digital Equipment Corporation
JAMES GRAY, Digital Equipment Corporation
JOHN L. HENNESSY, Stanford University
MITCHELL D. KAPOR, Electronic Frontier Foundation
SIDNEY KARIN, San Diego Supercomputer Center
RICHARD M. KARP, University of California at Berkeley
KEN KENNEDY, Rice University
ROBERT L. MARTIN, Bell Communications Research
ABRAHAM PELED, IBM T.J. Watson Research Center
WILLIAM PRESS, Harvard College
RAJ REDDY, Carnegie Mellon University
JEROME SALTZER, Massachusetts Institute of Technology
CHARLES L. SEITZ, California Institute of Technology
MARY SHAW, Carnegie Mellon University
EDWARD SHORTLIFFE, Stanford University School of Medicine
IVAN E. SUTHERLAND, Sun Microsystems
LAWRENCE T. TESLER, Apple Computer Inc.

MARJORY S. BLUMENTHAL, Director
HERBERT S. LIN, Senior Staff Officer
MONICA KRUEGER, Staff Officer
GREG MEDALIE, Staff Officer
FRANK PITTELLI, CSTB Consultant
RENEE A. HAWKINS, Staff Associate
DONNA F. ALLEN, Administrative Assistant
LESLIE WADE, Project Assistant

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND APPLICATIONS

RICHARD N. ZARE, Stanford University, *Chairman*
JOHN A. ARMSTRONG, IBM Corporation
PETER J. BICKEL, University of California at Berkeley
GEORGE F. CARRIER, Harvard University
GEORGE W. CLARK, Massachusetts Institute of Technology
MARYE ANNE FOX, University of Texas
AVNER FRIEDMAN, University of Minnesota
SUSAN L. GRAHAM, University of California at Berkeley
NEAL F. LANE, Rice University
ROBERT W. LUCKY, Bell Communications Research
CLAIRE E. MAX, Lawrence Livermore National Laboratory
CHRISTOPHER F. McKEE, University of California at Berkeley
JAMES W. MITCHELL, AT&T Bell Laboratories
RICHARD S. NICHOLSON, American Association for the Advancement of Science
ALAN SCHRIESHEIM, Argonne National Laboratory
A. RICHARD SEEBASS III, University of Colorado
KENNETH G. WILSON, Ohio State University

NORMAN METZGER, Executive Director

Preface

On December 13, 1991, at the request of the National Science Foundation, the Computer Science and Telecommunications Board (CSTB) of the National Research Council convened the Committee on a National Collaboratory: Establishing the User-Developer Partnership. The committee was charged to study and report on the need for and potential of information technology to support collaboration in the conduct of scientific research. An increasing number of scientific problems call for or would benefit from collaboration among researchers, while improvements in the capabilities, ease of use, and availability of computing and communications systems suggest that information technology can facilitate and enable collaboration.

The concept of a national collaboratory was first explored in a white paper written by William Wulf while he was assistant director of the National Science Foundation's Directorate for Computer and Information Science and Engineering.[1] Dr. Wulf coined the term "collaboratory" by combining the words "collaboration" and "laboratory." Initially proposed as a single all-encompassing entity, a national collaboratory was defined in the white paper as "a center without walls, in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries." The proposed collaboratory depended on a network of computers to perform its functions, but it was more than a mere interconnection of computers; it was envisioned as offering a complete infrastructure of software, hardware, and networked resources to enable a full range of collaborative work among scientists.

The national collaboratory concept was further explored in March 1989 at a 2-day workshop organized by biologist Joshua Lederberg and information technologist Keith Uncapher and held at the Rockefeller University. The workshop enthusiastically endorsed the concept and in general terms identified the technologies and services that would have to exist or be developed to provide the necessary infrastructure. The report of that workshop[2] and the diagram of collaboratory technologies (Appendix A) produced by Mark Stefik proved valuable resources for this committee.

This report is the result of a year-long effort to study the needs of scientists for computing and information technology to facilitate collaboration, and to relate those needs to the development and use of collaboratories.

The members of the committee met frequently during 1992 both in face-to-face discussions and by teleconference. To obtain information about scientists' requirements for collaboration and supporting technology and to get feedback on how collaboratories and information technology might facilitate research in the sciences, the committee held three 2-day workshops in specific disciplines: molecular

---

[1]Although Dr. Wulf became chair of CSTB in July 1992, he recused himself from CSTB oversight of this project and was not involved in discussions with the National Science Foundation about terms of the grant that supported this project. He served as assistant director for Computer and Information Science and Engineering from May 1988 to May 1990.

[2]*Towards a National Collaboratory*, the unpublished report of an invitational workshop held at the Rockefeller University, March 17-18, 1989 (Joshua Lederberg and Keith Uncapher, co-chairs).

biology (specifically genome research), oceanography, and space physics (Appendix B). These disciplines were chosen because they represent big and small science, a wide spectrum of technical and theoretical sophistication, and a broad range of institutions and styles of research, and they derive support from a variety of sources, including the National Science Foundation, the National Institutes of Health, the National Aeronautics and Space Administration, the (Defense) Advanced Research Projects Agency, the Office of Naval Research, and the Department of Energy. The workshops provided a rare opportunity for participants—scientists actively doing research in molecular biology, oceanography, and space physics; scientists and technologists specializing in computing, communications, and information science; and a sociologist specializing in the use of computing technology by scientists—to consider together a number of issues affecting the potential to build useful collaboratories.

As the study progressed, the idea of developing a single national collaboratory was replaced by the idea of developing multiple scientific collaboratories. These collaboratories would share network and computing resources, software, and infrastructure but would have unique features dictated by the needs of particular scientific disciplines. Recognition of those varying needs drove the shift in focus from a single national collaboratory to many scientific collaboratories.

Much of the report writing and editing was carried out using electronic mail on the Internet. The report's first four chapters relate the needs of scientists for information technology, generally to support collaboration and specifically to develop collaboratories. Chapters 2 and 3 address collaboration challenges and opportunities in oceanography and space physics, respectively. Chapter 4 examines a computationally intensive branch of molecular biology, genome research, and provides an information technologist's perspective on the building and designing of collaboratories to support genome studies. Chapter 5 provides a synthesis of the committee's observations on collaboratories, and Chapter 6 contains the committee's recommendations.

The time and talents of many people contributed to this report. The creative and thoughtful contributions of the workshop participants were essential to the committee's work. The committee thanks Tom Dickey of the University of Southern California for his substantial contributions to the report and the oceanography workshop, Bob Kahn of the Corporation for National Research Initiatives and Connie Pechura of the Institute of Medicine for their wise counsel, and the anonymous reviewers for their thought-provoking comments. The committee is also grateful to David Kingsbury of the George Washington University Medical Center; Peter Pearson of the Genome Data Base at Johns Hopkins University, who helped clarify the discussion of issues in genome research; John Wooley of the Department of Energy; and Chris Overton of the University of Pennsylvania. Of course, responsibility for the final content rests with the committee members.

Vinton G. Cerf, *Chairman*
Alastair G.W. Cameron, *Vice-Chairman*
Committee on a National Collaboratory:
Establishing the User-Developer Partnership

Contents

EXECUTIVE SUMMARY 1

1 SCIENCE, COLLABORATION, AND INFORMATION TECHNOLOGY 5
- Increase in Volume of Information and Complexity, 5
- Information Technology for Research and Collaboration, 6
- The Collaboratory Concept, 7
- This Study, 10
- Notes, 11

2 BUILDING COLLABORATORIES FOR OCEANOGRAPHY 12
- Oceanographic Research, 12
  - Field Experimentation, 13
  - Modeling, 16
- Collaborative Research in Oceanography, 17
  - International Initiatives, 17
    - World Ocean Circulation Experiment, 17
    - Tropical Ocean-Global Atmosphere Program, 17
  - Additional Interdisciplinary Programs, 20
- Using Collaboratory Components to Facilitate Research, 20
  - Improving Access to Colleagues, 21
    - Electronic Communication, 22
    - Educational Workshops, 23
  - Improving Access to Data, 24
  - Providing Tools for Collaboration in Oceanography, 25
    - TOGA Data Catalog, 25
    - Globe Data Catalog, 26
    - Tour Tool, 27
    - Cruise Planning Tool, 27
    - Ontology Tool, 28
- Attributes of a Useful Collaboratory for Oceanography, 28
  - Interoperability, 28
  - Transparency, 29
  - Customizability, 29
  - Integrity and Extensibility, 29
- Notes, 29

3 BUILDING COLLABORATORIES FOR SPACE PHYSICS 31
- Space Physics Research, 31
  - Data Collection and Instrumentation, 31

Methods and Technologies for Data Analysis, 32
Examples of Collaborative Efforts in Space Physics Research, 33
Initial and Ongoing Collaborative Programs, 33
Space Physics Analysis Network, 33
Coordinated Data Analysis Workshops, 34
Active Magnetospheric Particle Tracer Explorer/Charge Composition Explorer, 35
Sondre Stromfjord Observatory Testbed, 35
Solar-Terrestrial Energy Program, 36
Geospace Environment Modeling Program, 36
International Solar-Terrestrial Physics Program, 37
Space Physics Data System: A Collaboratory of One, 37
New Opportunities Provided by a Collaboratory for Space Physics, 38
Components of a Collaboratory Infrastructure, 39
Note, 40

4 BUILDING COLLABORATORIES FOR MOLECULAR BIOLOGY 41
Genome Research in Molecular Biology, 43
Trends in Technology to Support Collaborative Genome Research, 44
Current Database Systems, 44
MedLine, 45
GenBank, 45
Genome Data Base, 47
Future Information Systems—Toward a Working Collaboratory, 48
Basic Components and Research Prototypes, 49
Remote Analysis Servers—Blast, 49
Interconnecting Archival Databases—Entrez, 49
Community Information Systems, 51
Basic Elements and Capabilities, 51
A Model Collaboratory—Worm Community System, 52
Opportunities to Enhance Research, 53
Notes, 55

5 BUILDING AND USING NATIONAL COLLABORATORIES 56
Identifying Basic Capabilities a Collaboratory Should Support, 56
Providing Basic Capabilities—Technical Considerations, 57
Interconnecting Data Sources, 57
Sharing and Applying Programs, 58
Controlling Remote Instruments, 60
Supporting User Interaction, 60
Achieving Transparency, 62
Acknowledging Context—Social and Institutional Considerations, 63
Issues for Individual Scientists, 63
Costs for Individual Scientists of Using Computer-based Collaboration Technology, 65
Education and Training, 66
User-Developer Partnerships, 66
Issues for Individual Institutions, 68
Providing Local Infrastructure Support, 68
Managing the Results of Increased Interaction, 69

Issues in Funding for Collaboratory Infrastructure, 70
Making a Start, 71
Notes, 71

6 PROVIDING FOR A NATIONAL COLLABORATORY PROGRAM 73
Note, 77

REFERENCES 78

APPENDIXES

A Elements of a Functional Collaboratory, 83
B Workshop Programs and Participants, 86
C Rules Governing Access to and Use of the CDAW-9 Database, 95
D Training Computational and Mathematical Biologists, 97

Executive Summary

The fusion of computers and electronic communications has the potential to dramatically enhance the output and productivity of U.S. researchers. A major step toward realizing that potential can come from combining the interests of the scientific community at large with those of the computer science and engineering community to create integrated, tool-oriented computing and communications systems to support scientific collaboration. Such systems can be called "collaboratories."

Collaboration among colleagues is a challenge for the scientific community that takes many forms, most notably the sharing of data and/or special instruments, joint authoring of papers, and cooperative research. More and more scientific problems demand collaboration for their resolution as a consequence of increasing complexity and scale, a growing amount of which reflects the proliferation of fundamentally interdisciplinary problems. The study of global change phenomena illustrates all of these dimensions; it requires the expertise of oceanographers, meteorologists, biologists, chemists, physicists, experts in modeling and simulation, and others from around the world.

In many areas scientists have sought computer-based tools and techniques for data gathering, storage, analysis, modeling, and communication, making use of both generic (including off-the-shelf) technology and the tools they have developed to meet their own, specific needs. These bottom-up efforts have been productive, but their implementation has been difficult: funding for tool development has been inadequate, tools have been deemed awkward to use, and the building of tools is regarded by most scientists as less prestigious than the direct conduct of research.

At the same time (but largely in isolation), computer scientists and engineers have continued to advance the state of computer technology, developing better and less expensive tools for storing, accessing, and manipulating data; for monitoring and controlling instruments and other equipment; for supporting communications and collaboration among dispersed parties; and so on. But their general-purpose tools do not always match the needs of user communities. A more explicit partnership between scientists in general and computer scientists in particular can inspire development of computing technology that better meets the needs of scientists, better leverages the efforts of computer scientists, and provides broad benefits to scientists across the research community. That prospect was the motivation for this study, which focused on the potential for computer-based technology to facilitate scientific collaboration and improve the utility of computer-based resources used in scientific research.

Although technology will never cause the unwilling to collaborate, it can facilitate collaboration among those who are motivated and can also make it more attractive to others. There is evidence that this is happening. One example is the phenomenal growth in the provision and use of services offered through the Internet, the global network spawned by federally funded research into computer-based communications and now used by millions of scientists, engineers, and educators. Through the Internet, researchers access databases, share software and documents, and communicate with colleagues. The Internet has made collaboration among dispersed scientists practical, and it has been used for that purpose. Nevertheless, despite technological improvements, new tools, and guides, the Internet remains a somewhat primitive tool for collaboration, especially for those scientists who cannot enjoy or do not have the time for learning how to use it.

Observation of the rise of computational science and the popularity of the Internet and its constituent research networks among scientists in general led a group of computer scientists and engineers to conceive of and begin to explore the concept of a "collaboratory," which is an environment in which all of a scientist's instruments and information are virtually local, regardless of their actual locations. The virtual environment of the collaboratory supports interaction among scientists; among scientists, instruments, and data; and among networked computing tools used in the conduct of scientific research.

The development of a national collaboratory capability would facilitate collaboration of individuals and groups without regard to their physical locations. Collaboration may occur among scientists within a given facility or institution, but collaboration and interaction among distant researchers and resources are becoming increasingly important.

Although articulating the rationale for collaboration may be easy, achieving effective collaboration is not. In part, the situation reflects the basic training of scientists: scientists have been educated to focus on individual activity and achievement. Moreover, scientists have had to compete with each other to attain recognition and resources. Collaboration tends to be easier on a small scale and when it is local: when a small number of individuals collaborate it is generally possible to proceed on the basis of mutual trust, but "rules of the road" are needed for larger-scale collaboration. These and other human considerations shape and constrain the collaborations that do take place; in some instances they also inform the design of incentives to promote collaboration.

To gain insights into the motivations for collaboration among and within different fields of science, the obstacles to effective collaboration, and the potential benefits of computer-based tools for collaboration, the committee held workshops addressing these issues in the contexts of molecular biology, oceanography, and space physics, three fields that vary greatly in their use of computing and communications technology and in the applicability of the collaboratory concept. Despite these variations, all three fields share a common dependence on the collection and analysis of large amounts of data.

Through the workshops the committee found that collaboration is becoming more common (albeit at different rates) in these fields, within and between disciplines; that the conditions under which individual scientists work vary substantially; and that the familiarity with, access to, and use of computer-based technology vary significantly across fields. The workshops suggested that the broader community of researchers is aware of some of the relevant technological advances but often lacks the technical and financial support necessary for applying new technology.

The committee found that generally, any science that makes extensive use of computing for modeling, simulation, data analysis, and data storage and retrieval can benefit from the use of collaboratories, particularly in circumstances where collaboration has already begun. Bottom-up motivation will be an essential factor in the success of any collaboratory effort.

The committee concluded that a research program to further knowledge of how to implement and effectively use collaboratories would have broad impact. Such a program could involve the development, adaptation, or integration of wide-bandwidth communication between two or more sites allowing good transmission of sight and sound to achieve a virtual presence of an individual in someone else's laboratory, sets of database tools with common access and sophisticated graphics capabilities for interpreting masses of information, collaborative authoring and editing tools, and so on. While all of these developments can contribute to the conduct of science, the greatest impact will come from integrating these technologies and implementing them on a large enough scale to serve significant scientific communities—a large enough scale to provide scientists with new and better options for designing and executing their projects.

The committee envisions a program that would bring together computer scientists and other members of the scientific community. Such a program would present many challenges, given both the likely variations in the cultures of the disciplines involved and the potential awkwardness of having one partner in the position of a supplier and one in the position of customer. Nevertheless, the limited experience to date with such cross-disciplinary partnerships has been encouragingly beneficial to science.[1] Both the opening of all fields to more interdisciplinary activity and the recent, dramatic advances in

computing and communications technologies make the time propitious for joint collaboratory building by the scientific community.

RECOMMENDATIONS

The committee concluded that a collaboratory testbed program has the potential to address important scientific needs while simultaneously representing a key step toward developing national and global information infrastructure. The committee thus recommends an initiative to:

Establish a research program without delay to further knowledge of how to build, operate, and use collaboratories in the support of science. This program should have two major components of equal importance:

- **A *research component* dedicated to developing and integrating the software and hardware needed to build and apply collaboratories.**

- **An *education component* dedicated to educating and training the people needed to build and use collaboratories.**

The overarching goal of the program is to aid science and scientists through the construction and operation of working collaboratories. To achieve this goal the committee further recommends that the program:

Establish several collaboratory testbeds, funded at a level of $6 million per year each over a period of 5 years each.

Provide for two national demonstrations of each collaboratory testbed.

Initiate multiple and complementary activities to develop the human resources needed to carry out the collaboratory program, including, but not limited to:

A summer fellowship program to provide hands-on training for scientists and technologists in the use and development of collaboratory technologies in the conduct of science.

Regularly scheduled national symposia for testbed principal investigators, research staff, and graduate students, providing opportunities to share information, findings, and conclusions regarding the technical aspects of building, operating, and using collaboratories.

Because scientific disciplines vary in numerous ways, it should not be assumed that one mode of collaboration or one set of collaboration tools can fit all needs. Therefore, the committee recommends that the proposed research program establish several collaboratory testbeds, which should be tailored to the needs of particular groups of scientists in their respective fields. Although meeting specific needs is essential for collaboratories, developing multiple testbeds will allow an assessment of how much common infrastructure (such as the common support provided by the Internet) may be needed in comparison to more specialized tools. A 5-year, 3-testbed collaboratory program, for example, is estimated to cost about $100 million.

Achieving sufficient testbed scale is critical. The National Science Foundation's program on collaboration technology under the Directorate for Computer and Information Science and Engineering is already exploring research opportunities in this arena, but on a scale and scope less than the committee

believes necessary to create a national collaboratory capability. Demonstrating the revolutionary potential of collaboratories requires a new scale in the funding of research projects in information technology, because complete systems must be built, rather than isolated software, to demonstrate the usefulness of the full set of integrated technologies as well as the sociological effects. A national collaboratory can support a nationwide community interacting with a distributed digital library or with a remote large instrument. Implementing and operating such a system, even for an experimental testbed, requires an effort on the scale of national centers for databases or instruments. Hence, the major recommendation of this report is for the construction and evaluation of a few large collaboratory testbeds; the scale of the testbeds is more important than the number.

The introduction of a collaboratory testbed program must address explicitly a fundamental concern among the scientists who would be the users of collaboratories, namely the fear that spending on tools, regardless of their value, would diminish available resources for basic research in individual disciplines. This reaction may reflect past experience, inasmuch as scientists' efforts to build their own tools may have drawn on the funding and other resources nominally allocated for research. The proposed collaboratory program has been formulated to overcome this concern. First, it anticipates that as computation and communication become more integral to science, the distinction between "research" and "tools" will become less meaningful, and spending choices may appear less zero-sum. This is already happening in some areas (witness the broader attention to computational science). Second, by providing scientists with an alternative to do-it-yourself tool building, the program would help scientists (and their funders) to focus on their primary research while helping to assure that tools are built more efficiently and effectively. Meanwhile, the concept of a partnership between computer scientists and other scientists recognizes that there is mutual research benefit to the development of collaboration technology.

The proposed collaboratory testbed program is consonant with the objectives of the High Performance Computing and Communications (HPCC) initiative, which has the mission of developing high-performance computing systems, advanced software and algorithms, computer networking for research and education, and basic research programs and the human resources to support the mission. Many of the technologies developed as a result of HPCC will play important roles in the development of collaboratories. At the same time many components of HPCC, such as research into medical computation and environmental computation, access to academic medical centers and environmental databases, prototyping of experimental high-performance computing systems, research into software tools, and computer access in general will require collaboration between scientists and technologists. Finally, HPCC involves the (Defense) Advanced Research Projects Agency, Department of Energy, National Aeronautics and Space Administration, and National Institutes of Health as well as NSF, suggesting a broad potential base in government for relevant support and funding.

The committee believes that the program outlined by these recommendations is the minimum level of effort that should be undertaken. The research and education components, taken together as a program, constitute a strategic effort to improve the information infrastructure supporting the conduct of computationally intensive science in the United States.

NOTE

1. Several successful collaborations among natural scientists and their computer science counterparts offer powerful evidence of the potential of such work: the Heuristic DENDRAL and CONGEN projects; the Stanford SUMEX-AIM project; and the National Center for Biotechnology Information at the National Library of Medicine. (See, for example, Lederberg, 1978; and Carhart et al., 1975.)

1

Science, Collaboration, and Information Technology

Science is changing in many ways. The specifics vary among disciplines and subdisciplines, but, in general, scientists are addressing increasingly complex problems, the instruments and facilities needed to conduct research are becoming increasingly expensive,[1] and funding for scientists is becoming tighter, especially on a per capita basis. These are among the factors that are promoting broader interest in collaboration, although the nature and extent of collaboration differ across disciplines. They are also promoting broader interest in the efficiency of the scientific process, something to which more widespread use of computing and communications may contribute.

The scientific process encompasses a wide range of technical, social, and procedural activities (see Figure 1.1 for one view), each of which involves information—information is collected, combined, analyzed, derived, discussed, and distributed. Some, if not all, of these activities may and often do benefit from the application of computer and networking technology. All are subject to the knowledge, attitudes, and preferences of individual scientists, the prevailing norms or cultures in individual fields, and the constraints and incentives imposed by relevant institutions, from those that employ or educate researchers to those that fund or otherwise support them.

INCREASE IN VOLUME OF INFORMATION AND COMPLEXITY

Early in this century, the "gentleman's agreement" in astronomy was that when an individual began to publish observations of a certain class of stars, other astronomers would avoid observing that particular class. Now there are many more astronomers than there are classes of stars, and all types of observations are fair game for those who have suitable instrumentation at their disposal. Scientific activity has been growing so rapidly that the doubling time for the body of scientific information is now about 12 years, and today, at least 90 percent of all scientists who have ever lived are still alive.

Meanwhile, the nature of data collection has changed significantly, and the opportunities for making measurements have increased dramatically. For example, the ability to field automated instruments that can monitor conditions in distant reaches of space or in remote ocean regions has produced massive amounts of data. Moreover, new fields of science have arisen because it became possible to make particular kinds of observations.[2]

A result of the explosive growth in all the sciences is the sheer volume of information to be accessed, stored, analyzed, and understood. Any one individual can master only tiny fractions of the total of scientific knowledge. This is the individual who "knows more and more about less and less."

At the same time, the complexity of many of the problems now being addressed by scientists has led to an increase in interdisciplinary research, as well as to the recognition that computers and communications, or information technology, have become essential tools for handling complexity. The requirements for collecting and sharing massive amounts of data, the difficulties of developing and working with models of complex phenomena, the requirements for massive computation (achievable through shared high-performance computing systems or through interconnected distributed computing

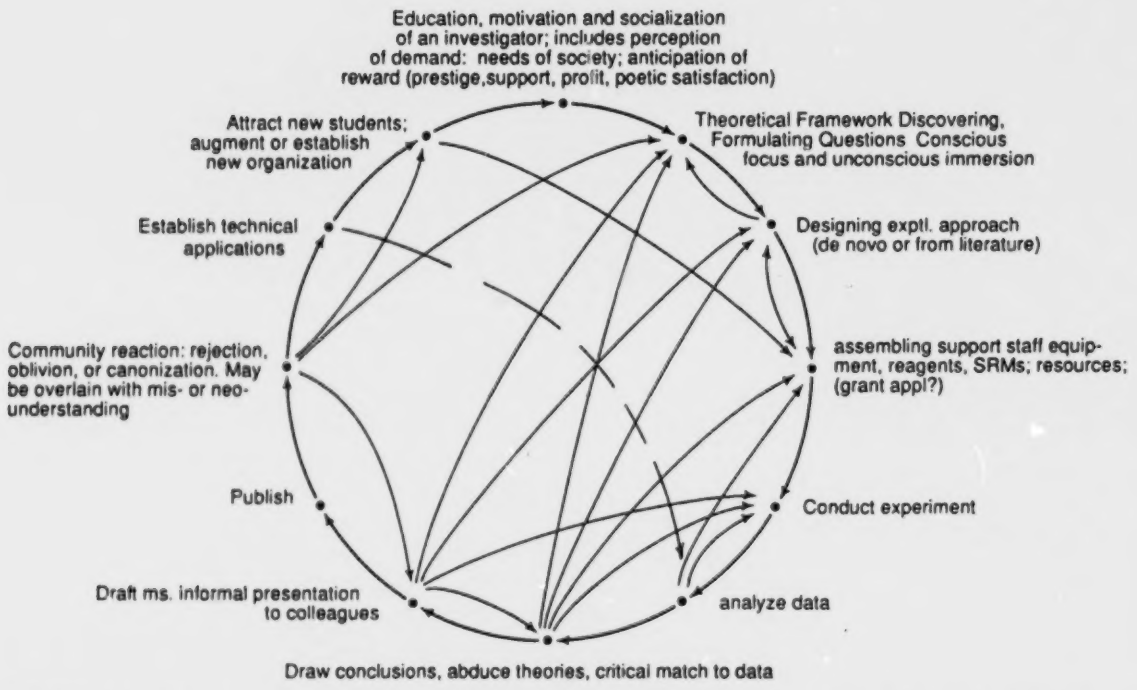

Education, motivation and socialization of an investigator; includes perception of demand: needs of society; anticipation of reward (prestige,support, profit, poetic satisfaction)
Theoretical Framework Discovering, Formulating Questions Conscious focus and unconscious immersion
Designing exptl. approach (de novo or from literature)
assembling support staff equipment, reagents, SRMs; resources; (grant appl?)
Conduct experiment
analyze data
Draw conclusions, abduce theories, critical match to data
Draft ms. informal presentation to colleagues
Publish
Community reaction: rejection, oblivion, or canonization. May be overlain with mis- or neo-understanding
Establish technical applications
Attract new students; augment or establish new organization

FIGURE 1.1 Epicycles of scientific discovery. The scientific process involves the development of theoretical frameworks and hypotheses, the testing of those frameworks through observation or experimentation, the analysis of data, the publication and discussion of results, and the education of people who will become scientists. Each of these activities is a complex process subject to significant uncertainties. Each activity can be undertaken in different ways and with differing outcomes, shaped by human nature and the prevailing conventional wisdom as well as by the history, maturity, and distinguishing characteristics of different scientific disciplines. SOURCE: Reprinted, with permission, from Joshua Lederberg's preface to *The Excitement and Fascination of Science*, Volume 3, Part 1, (c) 1990 by Annual Reviews, Inc.

systems), and the growing understanding that many pressing scientific problems transcend the boundaries of individual disciplines are other manifestations of complexity that drive demand for information technology in science. These conditions suggest that interdisciplinary research and collaboration, and the means for facilitating both, will become increasingly important to many scientists. It is often the case that the individual who "knows less and less about more and more" is needed to bring together individuals who can collaborate in correlating facts across disciplines.

Notwithstanding the growth in volume and complexity of scientific activity, funding for research is getting tighter. That context makes stretching and leveraging available dollars a necessity.

INFORMATION TECHNOLOGY FOR RESEARCH AND COLLABORATION

Differences among disciplines explain in part observed differences in the use of information technology and also in the propensity to collaborate and the styles of collaboration chosen. For disciplines that are data-driven, databases, libraries, and access to such resources are central requirements. Just to display (or "visualize") subsets of data is often a major challenge that calls for sophisticated algorithms and software as well as high-performance computing hardware. For disciplines that are more model-driven, algorithms and software are also key resources. Computer communications and related

information technology can also facilitate the automatic control and sharing of instrumentation, some of which may be local to a particular research endeavor and some of which is remote.

Many scientific projects have to devote substantial time and money to the mechanics of sharing resources and coordinating activity, time, and money that could better be spent doing science if the mechanics were easier. Many scientists are deeply frustrated with a computing environment that does not adequately support the demands of their science. Hardware problems range from difficulty in justifying purchases of computing equipment in grant applications to a lack of wiring in buildings to support local area network connections. Software problems range from an inability to access previously collected data to a lack of software for data analysis. Human resources problems range from not being able to find knowledgeable technical staff to not being able to pay scientifically attuned support personnel.

Moreover, despite the importance of collaboration, when too many human minds try to collaborate meaningfully, the requirements for communication become overwhelming. Facilitating the necessary robust communication among scientists involves both technical and social considerations—researchers must have access to useful computer facilities, networks, and data sets but must also be able to work in an environment that fosters cooperation among individuals with differing academic traditions, approaches to and priorities in research, and budget constraints (NRC, 1990a). Thus, it becomes necessary to choose the kinds of collaboration and computational aids that will enable the sharing of information, instruments, and ideas needed for science to advance so that understanding of phenomena is increased and practical benefits achieved.

There are now a number of national and international initiatives whose success depends on inter- and intra-disciplinary research and collaboration among scientists, including the U.S. Global Change Research Program, the World Climate Research Program, and the International Geosphere-Biosphere Program (NRC, 1990d). At the same time, the national High Performance Computing and Communications (HPCC) initiative promises to aid interdisciplinary groups of scientists, engineers, and mathematicians in applying emerging high-performance computing and communications systems to advance the solution of diverse science and engineering problems (FCCSET, 1992; OSTP, 1992).

For large interdisciplinary scientific initiatives, collaboration is a requirement of progress. It may work well or imperfectly, depending on a variety of personal and social factors (Box 1.1) as well as on the availability of appropriate technical infrastructure; lack of tools to facilitate communication, for instance, can impede the best-intentioned plans for joint work.

THE COLLABORATORY CONCEPT

The idea of using computing and networking technology to aid sciences other than computer science is not new (Box 1.2). Databases, electronic mail, and computer-based statistical packages for data analysis are all common tools across the sciences. What is new is the idea that various tools and technologies can be integrated to provide an environment that enables scientists to make more efficient use of scientific resources wherever they are located. Such an environment is termed a "collaboratory" in this report. At the highest level of abstraction, a collaboratory is a vision of a future ". . . 'center without walls,' in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries" (Wulf, 1989). In operational terms a collaboratory is a distributed computer system with networked laboratory instruments and data-gathering platforms; tools that enable a variety of collaborative activities; financial and human resources for maintaining, evolving, and assisting in the use of computer-based facilities; and digital libraries that include tools for organizing, describing, and managing data, thus enabling the large-scale sharing of data. A collaboratory provides a technological base specifically created to support interaction among scientists, instruments, and data networked to facilitate research conducted independent of distance.

BOX 1.1 SHARING AND COLLABORATION IN SCIENCE

Science advances through the process of sharing data, theories, ideas, and results. The scientific paper, published in peer reviewed journals or proceedings, is the preeminent formal mechanism for advancing collective understanding. Formal gatherings such as invited colloquia, conferences, and workshops are also important, as are informal exchanges among peers. In all cases shared information is scrutinized and critiqued. Meritorious work becomes part of the collective understanding, thereby advancing the collective enterprise.

Individual scientists, as well as the scientific enterprise, also advance through sharing, with the character of this sharing looking different within the primary work group from the way it looks outside that group. Within their primary work group, scientists engage in extensive sharing on a daily basis. Problem solving, invention, and interpretation occur in large measure through informal give and take, typically in face-to-face encounters. Two or three scientists gathered in heated discussion around an instrument display, data plot, or blackboard typify this kind of sharing, which is governed by norms of trust, reciprocity, and confidentiality. Through this mode of sharing scientists make sense of their own work and shape it for formal disclosure. Outside their primary work group, scientists' sharing occurs primarily through more stylized presentations and publications, which are governed by established norms and stand as accounts of work. Through this mode of sharing scientists receive credit for their work. Credit gets translated into reputation and other resources such as tenure, new and larger grants, better doctoral students, and scientific prizes. Because these resources are scarce, competition to secure credit can conflict with at least short-term open sharing. The need for timely publication and careful publication both militate against prepublication sharing of data with anyone outside a scientist's own work group. Strategic publication militates against publishing data (rather than papers) and publishing papers based on old data (rather than new data). In sum, the character of scientific sharing of data, theories, ideas, and results is influenced by both cooperation and competition.

Between collaboration and competition lies cooperation. Both collaboration and cooperation imply sharing data and other scientific resources. But the motivations and expected benefits are quite different. Cooperation may be impelled primarily out of narrow self-interest and may yield mutual benefit but not joint benefit. It can be construed as an exchange relationship. For instance, scientists cooperate with their peers by making their data available to them in publicly accessible databases. But they may do so primarily because they are required by third parties to make data public before they can publish or receive additional funding. Collaboration can be construed as a communal relationship that implies social trust and synergy among participants, with mutual benefit as the result.

Scientific organizations, like individual scientists, also engage in competition and collaboration. The competition for resources (funds and the best scientists) among scientific organizations is well known and intense. Yet, at the same time, cooperative agreements of different forms among organizations are also quite common today: academic and industry consortia, precompetitive industry projects, National Science Foundation science and technology centers, accelerator projects, and so on. While scientific organizations are motivated to advance scientific knowledge through these collaborations, they are also motivated to ensure their own well-being. Thus, issues of priority claims and credit can be as important to organizations as they are to individual scientists. (Some of the most complex provisions governing interorganizational agreements have to do with ownership of products resulting from the cooperation.)

A goal of the collaboratory concept is to render irrelevant the actual location of equipment and instrumentation and to make possible the creation of virtual laboratories using networked facilities. One can imagine the possibility of coordinating the capture of data by equipment on an orbiting satellite with the collection of data by ground-based instrumentation using computer networking tools to link all the facilities together. This collaboratory function may prove to be at least as important as providing for the sharing of information and support for collaborative interaction among colleagues. Where unique instru-

BOX 1.2 ANTECEDENTS OF THE COLLABORATORY

The antecedents of the collaboratory date to the development of the Arpanet in 1969. One of the first examples of computer-network-supported collaboration, started in 1973, was a collaboration among Stanford University, University College in London, and Bolt Beranek, and Newman in Boston. Using the Arpanet electronic mail and support for interprocess communication, a small number of implementors collaborated on the development of the TCP/IP protocols, the central element of Internet technology. By the mid to late 1970s, roughly 150 people linked by electronic mail were involved in developing the evolving TCP/IP suite (then numbering about 35 protocols). In the late 1980s, the Arpanet was decommissioned, superseded by the Internet, of which it had become an element. Currently about 1,500 people—scattered across 100 countries that cross all 24 time zones—are involved in some 80 working groups that constitute the Internet Engineering Task Force. Electronic mail, shared document databases, distribution lists, anonymous file transfer archives, and a cornucopia of new applications for distributed data recovery and management make this global collaboration feasible. These tools also enable a six-person staff to function as a secretariat for this rapidly paced technical standardization work. Industry participation in the work has led to the rapid development and deployment, sometimes within days, of products resulting from standards agreements.

Consisting of over 10,000 networks that link more than 1,500,000 computers, the Internet itself represents another kind of collaboration infrastructure. There is no central operating authority, and the system is funded by an international melange of private, public, for-profit, and non-profit resources. The system is doubling in size annually; it has spawned a multi-billion-dollar international equipment and service market in computer communications, and it is used worldwide for sharing scientific results and coordinating scientific research.

On-line publications are beginning to emerge from the Internet environment. The *Internet Society News*, for example, is published quarterly and incorporates the contributions of over 150 reporters worldwide who submit stories to the editor and to a page-layout editor over the network.

Many software development projects sponsored by (D)ARPA rely heavily on the Internet for project management and for collaboration. Common Lisp, a popular computer language for artificial intelligence, was developed by more than 60 people from universities, government, and industry who collaborated for 3 years but attended face-to-face meetings for only 2 days. According to a lead participant in the design effort, "The development of Common LISP would probably not have been possible without the electronic message system provided by the Arpanet. . . . Over the course of 30 months, approximately 3,000 messages were sent (an average of 3 per day), ranging in length from one line to 20 pages. . . . It would have been substantially more difficult to have conducted this discussion by any other means and would have required much more time" (Steele, 1984).

Another highly successful collaboration is exemplified by design activities using the Metal Oxide Semiconductor Implementation System (MOSIS). Developed by (D)ARPA in the late 1980s and early 1990s, this system first appeared as a prototype at Xerox PARC and was further developed by the Information Sciences Institute at the University of Southern California. It accepts very large scale integrated (VLSI) circuit designs in digital form over the Internet, combines multiple circuits where there is room on chips, produces a tape that describes the fabrication mask, arranges for fabrication of the wafers at a foundry, has the chips packaged and tested, and returns the circuits to the original designers within a few weeks at a cost ranging from a few hundred to a few thousand dollars per chip. This collaboration between circuit designers (principally, graduate students at U.S. universities) and the MOSIS project staff led to a significant increase in the number of trained VLSI designers and to the formation of some of the most successful computer hardware technologies, including the Geometry Engine used in machines produced by Silicon Graphics Inc. (SGI), systolic arrays that led to the Intel iWarp, the Connection Machine technology originally used by Thinking Machines Inc., and reduced instruction set computing (RISC) technologies such as those developed by MIPS Inc. (now part of SGI). The developing network and its collaboration-supporting electronic mail, file storage and retrieval capability, and the standards for chip design representation were all vital to the success of the effort.

ments are needed, or simultaneous data collection is necessary, a collaboratory's ability to manage and control local and remote instrumentation could easily make the difference between a successful experiment and an impossible dream.

In the information-rich world of scientific research today, discovery of relevant data and results is a major challenge. The information cataloging and indexing capability of digital libraries, which are part of the collaboratory concept, as well as the idea of having available the full content of reports and even the raw data and analysis programs used to process them, contributes to the appeal of collaboratories. With the proper information technology infrastructure, collaboratories could be formed quickly and flexibly to address particular problems or research opportunities.

Although variations may evolve over time and in response to the needs of different disciplines, a collaboratory may be envisioned as including up to perhaps 2,000 principal investigators, postdoctoral associates and doctoral students, scientific support personnel, and technical support personnel located at from 5 to 20 home institutions. Participants engaged in joint scientific research would be linked in a system providing computerized information technology for the collection, analysis, and distribution of data and results. All data or data products, and means of accessing instrumentation as well as all analysis and modeling capabilities and results, would be immediately available at every scientist's workstation.

To achieve a collaboratory capability, a considerable amount of research, development, and experimentation is needed. Although some features of a collaboratory (e.g., the types of instruments used, if any, the nature of the data collected, and the programs needed to analyze them) will be unique to particular disciplines, others will be more generally applicable across the sciences or even to meet more general needs for collaboration that are likely to emerge in commercial markets. Developing useful collaboratories thus requires research and development partnerships among scientists and information technologists to define, refine, and stabilize disciplinary or interdisciplinary collaboratory tools.

THIS STUDY

To further explore the concept of a collaboratory as it was first articulated and discussed in 1989 (*Towards a National Collaboratory,* 1989), the Computer Science and Telecommunications Board of the National Research Council convened a committee in December 1991 to study the need for and benefits of collaboration in scientific research, factors determining the effectiveness of collaboration, and the ability of information technology—specifically of electronically integrated collaboratories—to support and enhance interactive scientific research. In addressing these issues, the committee focused on three discrete areas of scientific investigation—oceanography, in which the difficulty and expense of gathering data and the interdependence of modelers and experimentalists provide motivation for greater collaboration (Chapter 2); space physics, which has of necessity used extensive computational technology in the analysis of data collected by cooperatively fielded space- and ground-based instruments (Chapter 3); and gene mapping and sequencing, research that has led to construction of and reliance on massive databases (Chapter 4). Research in these fields is sponsored by a variety of agencies, including the National Science Foundation, the National Institutes of Health, the National Aeronautics and Space Administration, the (Defense) Advanced Research Projects Agency, the Office of Naval Research, and the Department of Energy.

The committee's investigations suggested technical requirements and social and practical issues (Chapter 5) that must be considered and dealt with as part of the process of initiating a national collaboratory program (Chapter 6) in support of scientific research.

In conducting this study, the committee sought to:

- Identify common information technology needs that cross disciplines;
- Identify specific information technology needs in three particular fields of science, using this information to synthesize and refine the collaboratory concept;

- Increase awareness of the utility of information technology for the conduct of scientific research, particularly in the form of collaboratories; and
- Identify goals, objectives, and costs of developing collaboratories that would achieve concrete payoff in the form of enhanced scientific output.

NOTES

1. Space physicists, for example, use "smart" data collection instruments that incorporate microprocessors and the ability to discriminate among the data collected and thus to exclude unwanted "background" events. In several fields the ability to make important advances depends on having access to such instruments.

2. For example, recent declassification of truly revolutionary detectors developed by the military has invigorated the field of infrared astronomy.

2

Building Collaboratories for Oceanography

Research in oceanography covers a wide area and encompasses the marine aspects of several disciplines, including the physics, chemistry, biology, and geology of the ocean, the air-sea interface, the ocean bottom, and the shorelines. Yet the U.S. oceanographic community is relatively small, comprising about 4,500 oceanographers and ocean scientists in academia and in government laboratories. For them, the oceans continue to be a challenging environment in which to collect data, since the corrosive properties and hydrostatic pressure of seawater, mechanical failures associated with the forces of surface waves and currents, and the remoteness of many locations must all be dealt with. Considerable resources, relative to the available funding, must be expended to make the in situ observations required to advance the science. To an extent, an active modeling component of the community offsets the scarcity of and difficulty in obtaining field data.

Current and future capabilities that support electronic collaboration and improve access to data would do much to ameliorate some of the problems that impede the conduct of ocean science and would yield benefits realized in and beyond the oceanographic community. The ocean is known to be a large reservoir of heat, of carbon dioxide, and of other chemical constituents, and improved understanding of its role in weather and global climate change will contribute, for example, to enhanced capabilities for forecasting meteorological and longer-term climate conditions. To facilitate research in oceanography, to broaden the approach to addressing complex, large-scale processes that have significant social effects, and to carry on the learning associated with the larger, increasingly more cooperative process-oriented research studies carried out by the oceanographic community, a new thrust to observe the ocean on a global scale coupled with improved means of collaboration will be needed. Among such aids to collaboration are tools that can more effectively link members of the small, diverse oceanographic research community, both within and across the subdisciplines, that can provide more timely access to data collected at sea, and that can display and analyze information in ways that allow closer interaction between modelers and field experimentalists.

This chapter briefly describes oceanography and the present state of oceanographic research in general, outlines particular collaborative research programs, discusses the potential for improving computational support for current and collaborative research in oceanography, suggests the kinds of computer-based tools that would answer research needs of oceanographers, and lists the attributes of a useful collaboratory for oceanographic research.

OCEANOGRAPHIC RESEARCH

The U.S. oceanographic community includes physical oceanographers who study the dynamics and kinematics of fluid flows, including ocean currents and waves, and the forces that drive them; chemical oceanographers who focus on the distribution and variability of the ocean's chemical constituents; biological oceanographers who study the plants and animals, as individuals and as communities, found in the ocean; geological oceanographers who study sediments and rocks beneath the

oceans; and ocean engineers who design structures and instruments placed in the ocean and provide other related technology. Typically, academic oceanographers are either research staff or tenure-track faculty. For individuals in both groups, success in their careers depends on the ability to publish original, peer-reviewed research results in a timely fashion and at a regular rate. Within each oceanographic discipline, research efforts include field experimentation, numerical modeling, theory development, and laboratory experimentation. In field experiments, researchers have collected observations by working from ships and other crewed platforms and by deploying and leaving in place instruments that are moored, free-drifting, or placed on the sea bottom. More recently, observations of the ocean's surface have been made from satellites and aircraft.

Field Experimentation

In support of field work, the Universities National Oceanographic Laboratories (UNOLS) oversees the operation by academic institutions of research ships owned by the U.S. Navy, the National Science Foundation, or operating institutions. Available to investigators and scheduled roughly 1 year in advance, these ships make a series of research cruises all over the globe, returning to their home institutions perhaps once per year. The National Oceanic and Atmospheric Administration (NOAA) operates a fleet of about 20 ocean-going vessels, comparable in size to those of UNOLS, that are active in research, charting, marine resource assessment, and fisheries oceanography. Use of a UNOLS research ship costs approximately $15,000 per day, and 30-day cruises are typical.

The ships are used as platforms from which to make observations and to deploy instruments that are left in place on the ocean bottom, at the surface, or within the ocean (Figure 2.1). Instruments require considerable electrical power to operate and are designed to perform various tasks, such as collecting seawater, oceanic plants, or animals or making geographic surveys of ocean properties. Berthing space on research ships limits the number of scientists on board to roughly 12 to 24 individuals. However, because more than one investigator is on board and research must be carried on 24 hours per day, often only three to four scientists and technicians are available at one time to do the work associated with a specific project. At the same time, the equipment taken to sea has become more sophisticated, relying on advanced electronics and computers, and the need often arises for two-way ship-to-shore communication with engineers, technicians, and programmers unable to be on the ship due to space or financial limitations. Given the additional difficulties of staging experimental work from foreign ports, work done from ships remains a challenge.

Time series measurements are gathered by drifting and moored instruments (Figure 2.2) used to collect data sets that relate to specific events whose incidence is difficult to predict and match with the rather inflexible ship schedules. In addition, these instruments provide the capabilities of collecting unbroken data sets of up to several years in length and, when placed along a mooring line, of making observations at many different depths within the ocean. Oceanographic moorings are expensive, and their use requires a cruise both to deploy and recover the moorings. Recovery is a necessity not only to bring instruments back for reuse or maintenance but also to obtain the data they have collected. Cheaper, nonrecoverable instruments, called drifters because they are released to drift freely on the sea surface or within the ocean, deliver 100 to 200 data values per day using radio or satellite telemetry. When larger amounts of data are needed, instruments that store their own data must be used.

The limited data telemetry now done from moored and drifting instruments relies on links to satellites and usually is possible only when the instrumentation is located at the surface—as, for example, the meteorological sensors on surface buoys—or when oceanographic instruments are linked electromechanically or acoustically (using coded transmission of sound through the water to carry information) to surface packages. Radio frequencies do not penetrate seawater as they do air and space. Acoustic telemetry techniques, which are more suited for use in the ocean, are under development, but transmission of data through seawater is not routine. To work around this limitation, instruments have

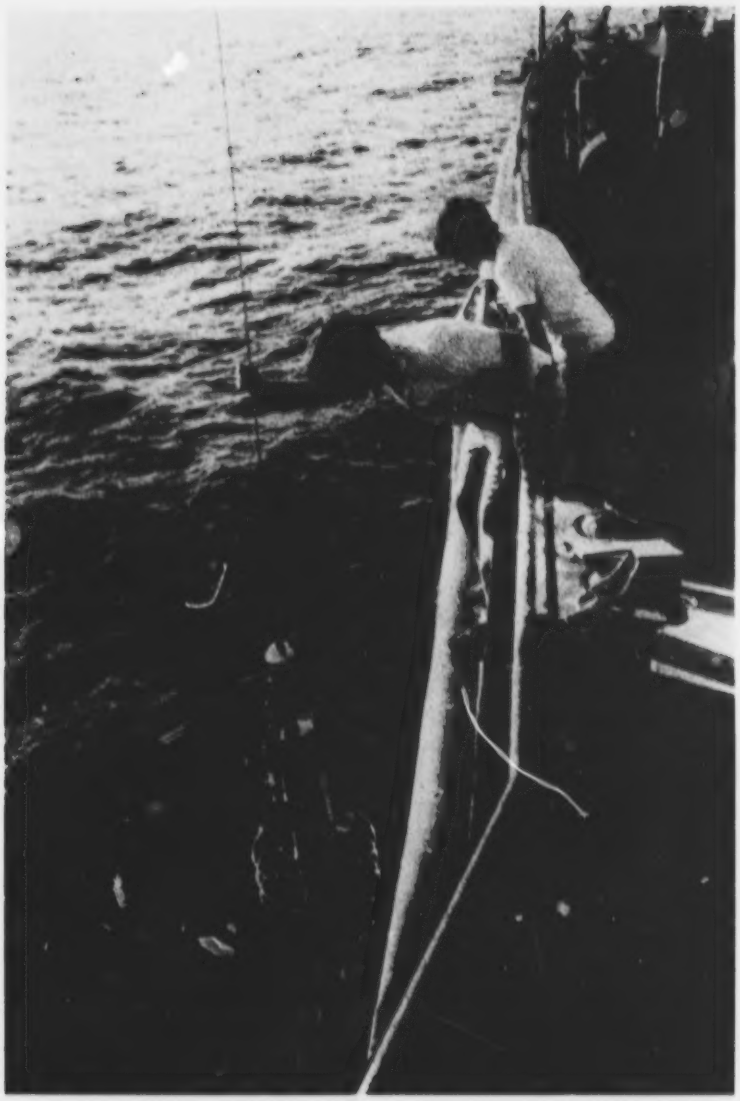
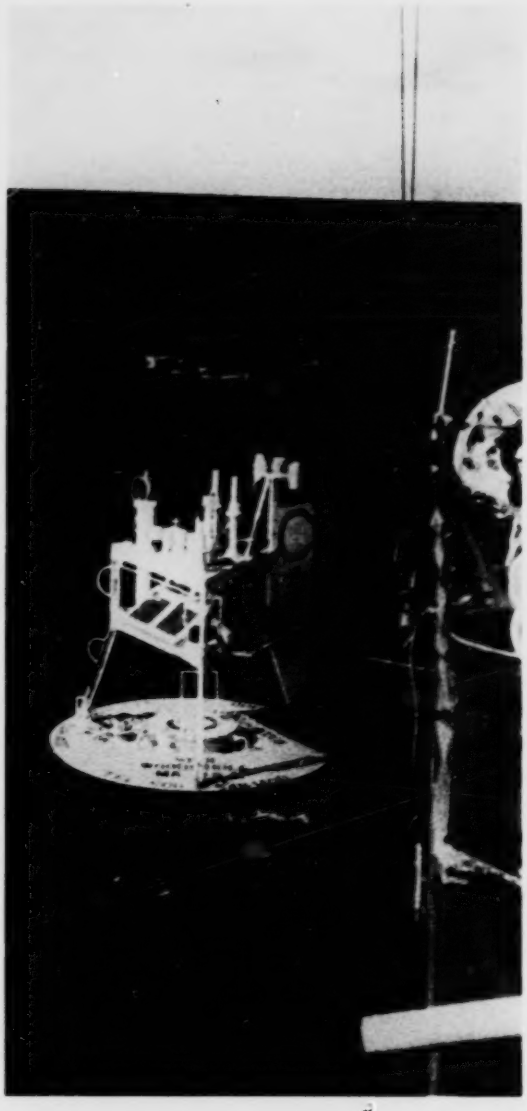


FIGURE 2.1 Conductivity, Temperature, Depth (CTD) oceanographic instrument being deployed from the side of the Research Vessel Oceanus, one of the UNOLS vessels operated by the Woods Hole Oceanographic Institution. Oceanographers use the CTD to collect profiles of the conductivity and temperature of seawater as a function of depth. Data from the instrument are carried by the conducting cable into the ship's laboratory. From measurements of conductivity and temperature, the density and salinity of the seawater are determined. Series of CTD stations along lines made during research cruises provide information about the distribution of the density of the seawater along a vertical plane or section. The large-scale ocean circulation perpendicular to the section can be calculated from the density data. SOURCE: Courtesy of R. Weller, Woods Hole Oceanographic Institution.

BEST COPY AVAILABLE



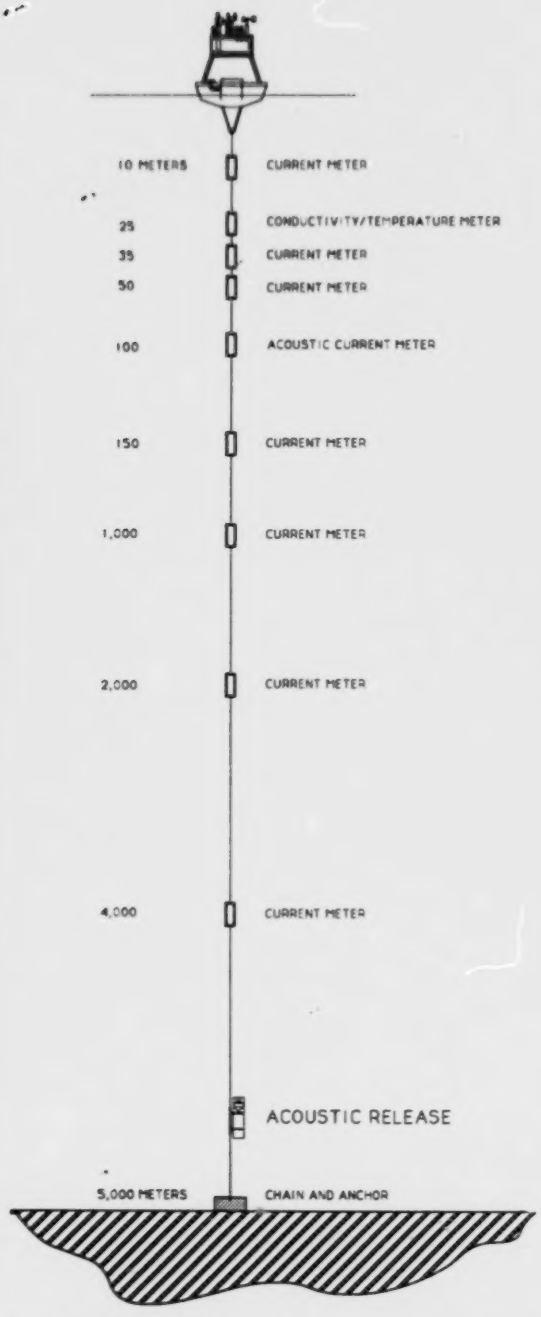


FIGURE 2.2 Photograph of a surface buoy and diagram of a surface mooring. Time series measurements of surface meteorology and oceanic parameters such as velocity, temperature, conductivity, dissolved oxygen content, and others are obtained by instruments left in place on oceanographic moorings. Sensors on the buoy being deployed from the research vessel measure wind velocity, incoming solar radiation, incoming long-wave radiation, barometric pressure, air temperature, sea surface temperature, precipitation, and relative humidity. The mooring line stretching between the buoy and the anchor will have oceanographic instruments attached as it is deployed from the ship. Some months later, the ship will return to recover the mooring and the instruments, which will have recorded data to magnetic tapes and disks. SOURCE: Courtesy of R. Weller, Woods Hole Oceanographic Institution.

BEST COPY AVAILABLE

BOX 2.1 DATA COMMUNICATION FROM OCEANOGRAPHIC INSTRUMENTS

Polar-orbiting satellites are often used for relaying oceanographic data. Transmitters mounted on buoys or ships for use with the Argos data collection system on polar-orbiting NOAA satellites transmit 32 bytes of data roughly every minute. However, the satellite passes overhead only 5 to 15 times a day and during any overpass is in view for only about 10 minutes. Because of the limited data rate possible with this system, transmission of a complete set of hourly meteorological data (wind velocity, barometric pressure, solar radiation, long-wave radiation, sea surface temperature, air temperature, relative humidity, and precipitation) requires use of multiple transmitters and transmission of the contents of a ring buffer (the data from 4 hours are stored and transmitted over and over again, but are updated as new data become available) from each transmitter. However, with the cost of use of Argos at about $4,000 per transmitter for users in countries participating in an Argos joint operating agreement, this method is not used often. (More information about oceanographic data telemetry is available in Dickey et al. (1993) and in Briscoe and Frye (1987).)

recently been designed that rise to the surface to access the satellites and transmit the data that have been gathered. However, transmission of data collected at sea by moored buoys, drifters, and ship is the exception and not the rule, due to limited transmission capacity and the great expense associated with the process (Box 2.1). The commonest way to collect data from instruments at sea is to undertake a cruise to retrieve the instruments.

In part because of the cost of maintaining and staffing research ships and moorings and the difficulty of achieving access to near-real-time data, no worldwide, comprehensive, operational oceanographic observing system, such as exists in the atmosphere to support weather prediction, is in place to monitor the variability of the oceans. Oceanographic field work to date has focused on specific hypotheses and deployed its limited resources for short periods to investigate specific processes. However, financial constraints and the inaccessibility of data are not the only reasons for lack of a global ocean observing system. Having seen much of their growth since World War II, the ocean sciences are relatively young and are characterized by having gathered comparatively few data from the ocean. Current field work, which is very productive scientifically and is contributing to a growing understanding of the time and space scales of the processes at work in the ocean, is thus essential to building the foundation of understanding required to intelligently plan observations on a global scale.[1]

Modeling

Modeling serves several purposes in the study of the oceans, among them (1) complementing sparse oceanographic data, (2) providing a means to test our understanding of the processes at work in the ocean, and (3) producing forecasts (there is great interest, for example, in developing the ability to predict the occurrence of El Niño). Modeling now under way addresses the ocean's role in climate change, the maintenance and variability of the large-scale ocean circulation, and the interaction of the ocean circulation with the atmosphere, the biosphere, the hydrological cycle, and the solid earth.

Modelers are found in small numbers (perhaps only one or two in some cases) at many institutions, although centers of activity exist at NOAA's Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey, at the National Center for Atmospheric Research in Boulder, Colorado, and at Navy laboratories. Ocean modelers in general require access to large databases with, for example, climatological data that specify the initial state of the ocean or the annual variability of the surface forcing of the ocean by the atmosphere; high-power workstations and access to supercomputers; specialized data-

display software, and the means to collaborate with colleagues at other institutions while writing proposals, reports, and publications. Modelers need to store, manipulate, exchange, and visualize the data produced by models; for a global model of the ocean with high temporal (every 4 hours over a period of many years) and spatial (every 100 km) resolution, the data files could be as large as 1 Gbyte.

Because of the complexity of ocean models (e.g., ranging from direct eddy simulation models of the small scale to global general circulation models) and the need for powerful computing resources, relatively few oceanographers have access to ocean models as research tools. However, a growing number of modelers are working with shared (community) models. Thus, the models are in a sense community resources and represent a form of collaboration that maximizes the use of precious resources for scientific advances.

COLLABORATIVE RESEARCH IN OCEANOGRAPHY

International Initiatives

World Ocean Circulation Experiment

As described in a recent report (NRC, 1990b, pp. 4-5), the World Ocean Circulation Experiment (WOCE), an international program that is part of the World Climate Research Programme,

> was created because an understanding of ocean circulation is crucial for predicting global climate change. Circulation is related to climate on a decades-to-centuries scale, through the transfer of heat and momentum between the atmosphere and the ocean. WOCE will study surface and subsurface circulation of the world's oceans over a seven-year period, with the goal of understanding circulation well enough to model its present state, to predict its future state under a variety of assumptions, and to predict feedbacks between climate change and ocean circulation. These goals would be met by describing (1) the ocean's present circulation and its variability, (2) air/sea boundary layer processes, (3) the role of exchange between different ocean basins in global circulation, and (4) the role of the oceanic heat storage and transport on the global heat balance.
>
> The WOCE program is divided into several interlocking parts, the largest of which is the global survey carried out in international cooperation. This global hydrographic survey will carry out a number of cross-ocean sections sampling (1) water density, which helps drive ocean circulation, (2) various natural tracers, such as oxygen and nutrients, and (3) man-made tracers of water motions, such as chlorofluorocarbons. Worldwide placement of floats and current meter moorings will augment the global survey with direct observations of current ocean velocity. Important objectives are quantifying the oceanic transport of heat and the pathways of downward movement of water by which atmospheric gases are transported into the deep ocean, and to correctly model observed circulation patterns. An upper ocean program will focus on the atmosphere-ocean fluxes that drive the ocean and feedback to the atmosphere and on variations of the upper ocean temperature and heat storage. Satellites, voluntary observing ships, moorings, and surface drifters will be integrated into an observation system capable of global measurements.

Tropical Ocean-Global Atmosphere Program

The Tropical Ocean-Global Atmosphere (TOGA) program is a decade-long international research program begun in 1985 to measure, model, understand, and predict variability in global climate associated with the El Niño-Southern Oscillation (ENSO) phenomenon (National Research Council, 1990c). ENSO is a robust, identifiable, recurrent (roughly every 4 to 7 years) climate signal whose origins can be traced to the tropical Pacific and whose impacts are felt worldwide through perturbations of the atmospheric

general circulation. The TOGA concept derives from a theory of coupled ocean-atmosphere interactions in the tropical Pacific first articulated by Jacob Bjerknes in the 1960s and subsequently expanded by other investigators; these studies provided impetus for the planning of TOGA in the early 1980s, in the midst of which the 1982-1983 ENSO event occurred. The 1982-1983 ENSO event was the most intense of the century, leaving in its wake human misery and billions of dollars in devastation on a global scale. Significantly, development of the 1982-1983 ENSO went undetected until the event was well under way. This dramatized to the scientific community the need to monitor the tropical oceans in real time to detect precursors of ENSO and to develop models capable of skillfully predicting ENSO events months to years in advance.

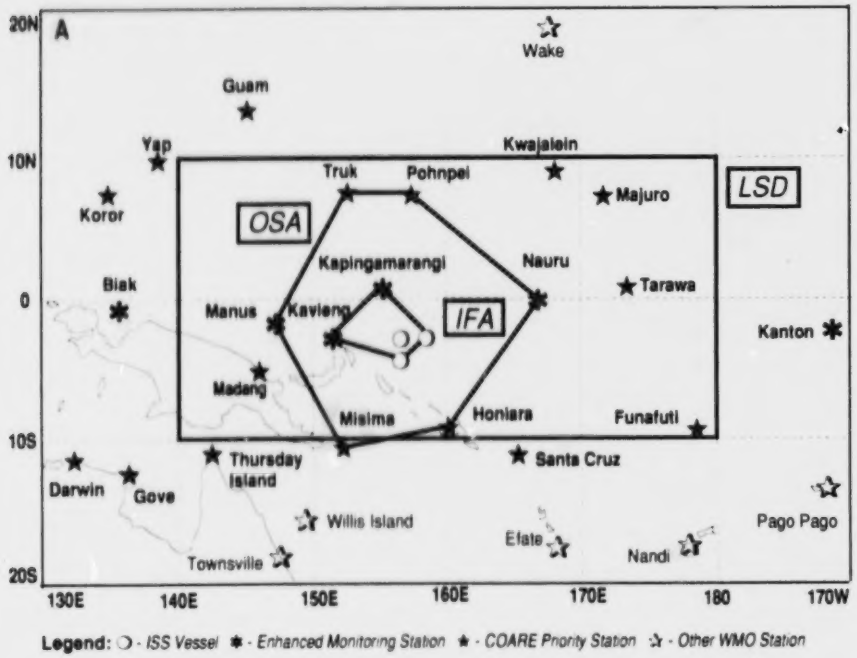
These two capabilities, real-time monitoring and model-based climate prediction, have guided the implementation of an ocean-observing array in support of TOGA objectives. TOGA observations fall into two broad categories, namely, in situ observations and remotely sensed satellite observations. Among the most valuable of the satellite data collected to date have been measurements of sea surface temperatures from NOAA weather satellites and of sea level from the U.S. Navy's Geosat mission. These data have been useful in defining patterns of surface variability on time and space scales relevant to understanding the ENSO phenomenon. They have also been valuable for assimilation into and/or validation of ocean models under development for climate prediction.

Given the reality of limited resources, TOGA has concentrated its in situ observational efforts in the areas of the Pacific Ocean where the scientific issues related to ENSO events are most clearly defined and, from a global perspective, are most compelling (Figure 2.3). Much of the data are transmitted in real time via polar-orbiting weather satellites (Satellite Service Argos) or geostationary satellites for use by the research community. Increasing amounts of in situ data are also being disseminated on the Global Telecommunications System to national meteorological centers for use in operational weather prediction.

One measure of progress during the first half of TOGA is the development of statistical, statistical-dynamical, and purely dynamical models that exhibit limited though significant skill for predicting ENSO events several months to a year in advance. This success has stimulated discussion not only of how to best capitalize during the second half of TOGA on the scientific advances that have been made, but also of how to best translate these advances into an operational system for climate prediction during the post-TOGA period (1995 and beyond).

TOGA is unique not only for its scientific contributions to understanding and predicting ENSO events and related phenomena, but also for the way the oceanographic community has organized itself to make these contributions. The compelling need for obtaining data in real time has led to a more collegial attitude among oceanographers toward the sharing of data for both operational and research purposes. In many instances, the customary 2-year period of exclusive or proprietary rights to the analysis of a new data set in oceanography has been waived by a particular investigator in the interests of furthering the common goal of improved understanding and prediction of short-term climate variability. This shift in attitude is by no means universal within the TOGA oceanographic community, nor has it been completely voluntary. It has been fostered in part by the peer review process, which has favored grants to investigators who propose to disseminate data in real time (or in near-real time) to a broad spectrum of investigators.

Another indication of the sociological transformation under way among oceanographers involved in TOGA is the gradual movement toward a modus operandi similar to that in meteorology and stemming from the nature of the scientific problems being addressed. In meteorology, much of the data collection effort is driven by the need for improved numerical weather prediction, and most observations for this purpose are supported by the intergovernmental World Weather Watch. A long tradition of operational support for meteorological observations obviates the need for most research meteorologists to become involved in field work. Oceanography, by comparison, is a relatively new field of study and, until recently, has not had a clearly defined operational imperative to support climate prediction. Hence much



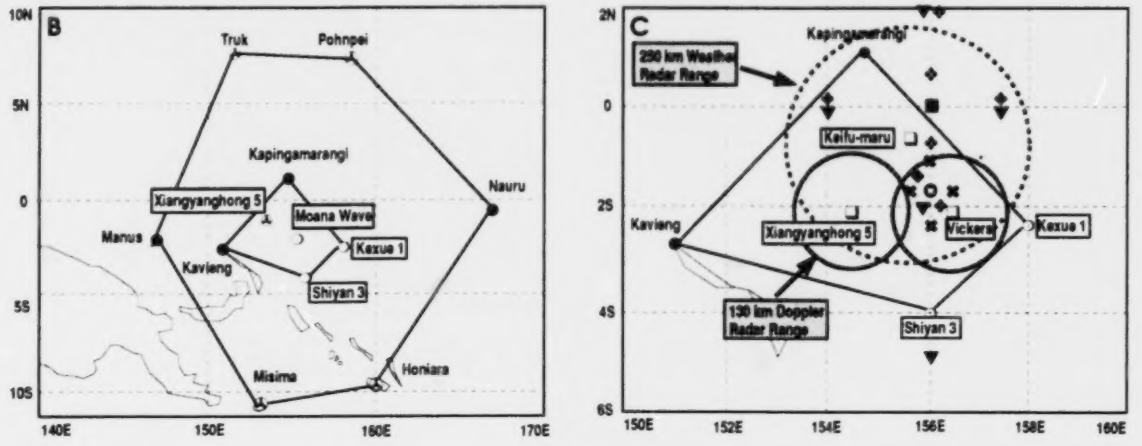


FIGURE 2.3 Embedded within TOGA, which covers the entire equatorial Pacific region, is the Coupled Ocean-Atmosphere Response Experiment (COARE), a major process study that reflects the complexity of modern process-oriented oceanographic research efforts. Shown is the structure of the intensive observation period of TOGA COARE. The legends beneath the panels define the symbols used to represent the observational platforms. (A) The entire COARE domain; the large-scale domain (LSD), the outer sounding array (OSA), and the intensive flux array (IFA) are outlined. (B) Location of the observational components of the OSA, including land-based and ship-board components. (C) Locations of ship-borne components relative to the many oceanographic moorings in the IFA. SOURCE: Reprinted, with permission, from Webster and Lukas (1992).

of the data used in oceanographic research has been, and still is, collected via the mechanism of individual peer-reviewed proposals. In view of the daunting logistical and technical challenges involved in oceanographic field work, investigators who successfully compete in the peer review process have traditionally had little incentive to share hard-earned, highly prized data sets too freely. In TOGA, on the other hand, real-time and near-real-time oceanographic data streams are voluminous and continue to grow. In parallel with the evolution of the TOGA ocean-observing array, numerical models and ocean data assimilation techniques suitable for climate studies have also undergone rapid development. It is now generally recognized that long-term maintenance of the TOGA ocean-observing array should become the responsibility of operational agencies, since the justifications for large-scale measurements are cast increasingly in terms of initializing and verifying operational ocean models for climate prediction. Accordingly, NOAA's National Ocean Service has become an ever more prominent source of support for TOGA observations, as in the case of the tide-gauge sea level network and the Ship of Opportunity Program/Expendable Bathythermograph Program.

Additional Interdisciplinary Programs

Interdisciplinary oceanographic research will be increasingly common, as exemplified by programs such as the Global Ecosystems Dynamics Experiment (GLOBEC) and the Joint Global Ocean Flux Study (JGOFS). GLOBEC is designed to evaluate how changes in global climate and related physical processes influence the ability of individual animals to feed, grow, reproduce, and survive in the sea. JGOFS focuses on the climatic implications of time-varying fluxes of greenhouse gases (e.g., carbon dioxide) as related to physical forcing, bringing together at sites near Bermuda and Hawaii scientists who examine biological, geochemical, and physical processes on time scales ranging from months to years.

Successfully realizing the goals of such programs will almost certainly require a concerted effort to ensure that the electronic infrastructure supports and sustains collaboration across the disciplines, whose tools and data types can differ greatly. In addition, the diversity of the data (Figure 2.4) collected in such interdisciplinary programs will itself present a challenge to collaboration. Some physical data are digitized in the instruments and made available soon after collection via satellite telemetry, whereas some biological data, such as population statistics, may not be available for exchange until months after completion of the cruises in which the sampling was done. It will be a challenge to establish a database with the life span and flexibility needed to serve all participants in the large interdisciplinary programs under discussion, which potentially could generate 25 Gbytes of data per year.

USING COLLABORATORY COMPONENTS TO FACILITATE RESEARCH

The oceanographic research community is widely distributed geographically, comprises academic, government, and private-sector scientists, and has many subdisciplines. Researchers in field programs commonly collaborate; single-investigator experiments are now rare. Increasingly, the high cost of field work has led to large, cooperative experiments in which investigators pool instrumentation and share time on ships. Thus investigators at diverse locations need ways to plan cooperatively, write coordinated proposals, communicate between different platforms in the field, and work together to analyze and publish their results.

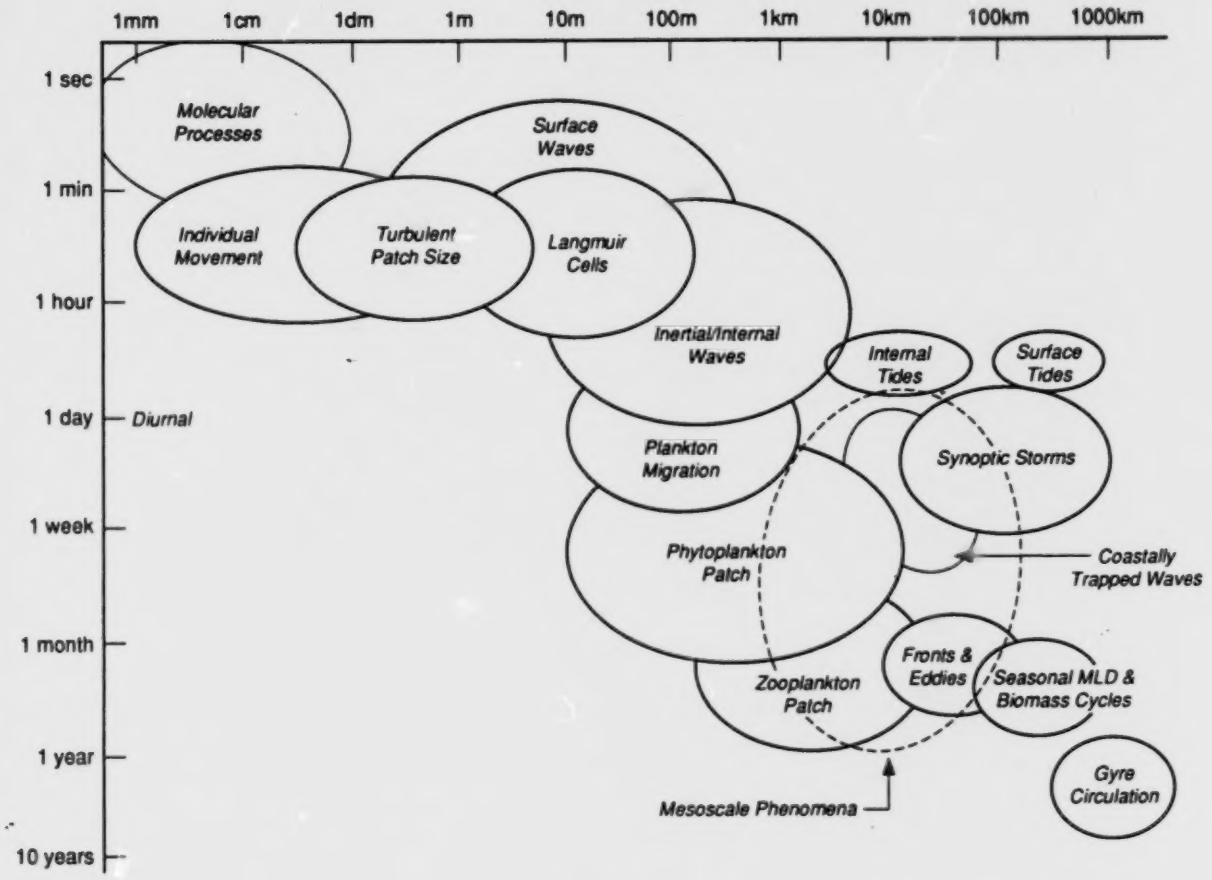


FIGURE 2.4 Schematic diagram showing the relevant time and space scales of several physical and biological processes that need to be considered in studies of the physics, biogeochemistry, and ecosystems of the upper ocean. SOURCE: Reprinted, with permission, from Dickey (1991).

Improving Access to Colleagues

Increasingly, the problems of current interest are either interdisciplinary or global, or both. As a consequence, research programs not only cross the lines of the traditional subdisciplines of oceanography (physical, biological, chemical, and geological) but also require collaboration among international scientists. The need for ongoing dialogue with colleagues exists as well between theoreticians, numerical modelers, and laboratory modelers. Traditionally, much of the required interaction has been done at meetings, requiring participants to travel to a common location. Now, however, at the same time that more widespread collaboration is needed, travel funds and the time scientists have available to travel are both more difficult to come by. Ways to facilitate collaboration across disciplines and at a distance are needed.

BOX 2.2 BRIEF HISTORY AND DESCRIPTION OF OMNET INC.'S SCIENCENET

In 1979, in response to the need of participants in large research programs for a communications alternative superior to telex, real-time telephone communication, or voluminous photocopying, pilot electronic mail networks were set up for three oceanographic research programs and managed from the Massachusetts Institute of Technology. Omnet Inc., started in 1980, provides the on-line service SCIENCEnet. The first group to use SCIENCEnet was the NSF-funded Pacific Equatorial Ocean Dynamics Experiment (PEQUOD), which involved 40 program participants from the United States, Australia, and Canada. Over the next decade, the growing network included primarily the international earth sciences community.

Simple electronic mail provided program groups a means for communicating without working together in real time. Meetings were scheduled, agendas modified and clarified. Over time, SCIENCEnet expanded beyond simple communications, and participants found new ways to use the simple bulletin board structure. Notices were posted on subjects ranging from calls for papers and meeting announcements to job postings. De facto conferences sprang up. "ENSO.Info" (El Niño-Southern Oscillation Information), for instance, is a worldwide discussion of the likelihood of El Niño events occurring in the southern Pacific, the accuracy of various models, and related topics. Bulletin boards have been used to locate lost deep-sea research buoys. The "Gulf.Mex" board became a repository of maps of oceanographic data from the Gulf of Mexico, using an ASCII format devised by one of the participants. Joint documents were created. Schedules and calendars were shared.

SCIENCEnet has a library of custom electronic message forms that allow participants to submit annual reports to funding agencies such as the Office of Naval Research and the National Science Foundation, or to register for meetings. Under the pressure of individual, programmatic, and agency needs, SCIENCEnet evolved into an early, effective collaboratory working environment.

The early need to connect to the network from research vessels at sea was first accomplished in the early 1980s via a voice channel on a communications satellite, the Applications Technology Satellite (ATS), used by the ocean research community. As seagoing researchers became more dependent on such access and needed to move outside the footprint of the ATS satellite, much of the ship-to-shore traffic moved to a commercial satellite system, Inmarsat. SCIENCEnet now has participants in 50 countries and has connected researchers in remote parts of the world. For example, it currently provides communications for Antarctic research stations and a remote ice-coring party on the Greenland ice sheet. In 1989, SCIENCEnet provided a link for a climbing party on Mt. Everest to receive customized weather forecasts from the University of Pennsylvania.

SCIENCEnet does not provide computational resources, but users have not demanded that it do so, owing to the availability of cheap local computational power on individuals' desktops at one end of the computing spectrum, and access to national supercomputing centers at the other. (It is clear, though, that program source code has been a nonnegligible part of the message traffic.) Plans for the future include a closer integration of SCIENCEnet with the Internet, from which Omnet Inc. currently provides gateway access to SCIENCEnet, and with Joint Oceanographic Institutions Inc., the development of SeaNet, an extension of the Internet to ships at sea. Also, the capabilities of the network are being expanded to provide for multiple-author document preparation, a simple conferencing system, better directory services, and an improved database capability.

Electronic Communication

Electronic communication through SCIENCEnet is one example of how the oceanographic community has attempted to establish links and infrastructure that permit collaboration (Box 2.2). Run by Omnet Inc., SCIENCEnet provides a user-friendly environment, user support, and customized communication for its paying customers and is an important part of the communications infrastructure

for oceanography. However, there continues to be a pressing need for improved infrastructure within the oceanographic community.

Improvements to the communications infrastructure that boost connectivity among researchers and between researchers and remote platforms would have clear benefits. For example, improved voice and computer access during data-gathering cruises would allow more technical support to be carried out by land-based personnel. A project under development by Omnet Inc. and Joint Oceanographic Institutions Inc. is intended to extend the Internet to ships, buoys, aircraft, and other remote platforms that exist in the ocean environment. Called SeaNet, the project seeks to provide low-cost/high-bandwidth data transmission capabilities to oceanographers. Although it is still in the earliest stages of development, this project, should it succeed, promises to have a significant positive impact on the oceanographic community.

Collaboration tools that initially help modelers to work together, such as electronic mail, video links, and methods for exchange and common display of data, would be valuable, as would tools that build intuition about a process being modeled. Such tools would also have value beyond the modeling community. Modeling can guide the planning of field work, and tools that simplify access to and understanding of model results would make it easier for field experimentalists to gain from and interact with modeling efforts. In the future, parallel processing techniques, three-dimensional modeling and visualization methods, including movies and other animation techniques, and electronic publishing are anticipated.

One positive aspect of the developing electronic communications infrastructure and potential collaboratories for oceanography is the overt recognition of the need for computer and information specialists within oceanography and the establishment of exchanges between such specialists and the larger computer science and information community in the United States. This may particularly aid the developing oceanographic modeling community.

Educational Workshops

To make progress on the "grand questions" of oceanography, oceanographers, most of whom have been trained in a traditional, narrow subset of the field, must strive to see the ocean as a system and not just as being of isolated physical, biological, or chemical interest. New ways of pooling and applying the diverse skills of scientists whose formal training may vary considerably—in effect, of developing efficient collaborations—are thus needed. One approach is to bring faculty, graduate students, and principal investigators together to explore research issues in an extended workshop (Box 2.3). The components of a collaboratory workshop include:

- An interdisciplinary group of interested scientists (the core faculty) to teach the fundamental science of a particular project;
- Participants consisting of a combination of advanced graduate students and principal investigators;
- A period of 4 to 6 weeks in which to conduct the workshop;
- Funding by an interagency group whose mission is furthered by the workshop; and
- A computer network that provides access to databases and numerical models relevant to the scientific project.

Ideally, participation in such an intensive period of study stimulates graduate students to think about particular problems in a "systems" mode, establishes early in a scientific career lasting student interactions that cross the traditional subdisciplines of oceanography, and encourages principal investigators from diverse subdisciplines to collaborate immediately to share data, write papers, and propose new science.

BOX 2.3 OCEANOGRAPHIC MODELING—AN EDUCATIONAL WORKSHOP

The paradigm of a collaboratory workshop has been used by Lewis Rothstein of the University of Rhode Island, who held an educational workshop in June 1991 for participants in the Joint Global Ocean Flux Study (JGOFS). The focus of the workshop was the physics of the equatorial ocean; the goal was to define the fundamental physical processes critical to understanding biogeochemical cycling in the region. A group of 4 resident faculty, 8 guest faculty, 33 graduate students, and a number of principal investigators participated at the University of Rhode Island's Graduate School of Oceanography. A computer network established for the workshop consisted locally of 15 Unix graphics workstations, two color printers, video recording equipment, and a variety of peripheral devices networked to the CRAY Y/MP at the National Center for Atmospheric Research. Each student was allowed 10 hours of CRAY time for numerical experimentation, under the guidance of the faculty. Physical models of the equatorial circulation, as well as models of ecosystems, were prepared to provide students with a basis for asking "what if" questions of the "system": e.g., How would the ecosystem respond to a reduction in equatorial wind stress? The students were able to produce movies of the resulting simulations, which provided rapid graphical feedback to enhance the educational process. The principal investigators used the same tools to extend previous results, collaborate on new papers, and define potential future projects. Thus, the workshop furthered the short-term goals of the JGOFS program as well as the long-term educational goals of the students and faculty. Such an extended workshop is an attractive model for future efforts to provoke new ways of thinking about ocean system science.

Improving Access to Data

Oceanographic data are diverse in nature because of varied sampling platforms, the interdisciplinary problems considered, and the variety of required sampling techniques (e.g., optical, acoustical, and physical). Data take forms ranging from point time series to global spatial maps. Near-real-time data are needed for several purposes, including linking in situ and satellite data sets, conditional sampling, and modeling. More generally, models require satellite and in situ data sets for verification of both initial and boundary conditions. Often, however, one research program produces a mix of telemetered, near-real-time, and delayed data, with some data available soon after the end of the program and other data available only after several additional years of processing and analysis. Thus, access to and synthesis of data remain problematic for many oceanographers. Reliable, affordable data transmission from sea to land and policies that establish incentives and protection for sharing data are crucial components of an improved communications infrastructure for oceanography.

Particularly needed is development of a unified system for the collection, processing, and distribution of interdisciplinary, in situ data collected from ships, mooring buoys, drifters, and, in the future, autonomous underwater vehicles. The present Argos satellite data communication system (see Briscoe and Frye (1987) for a summary of current oceanographic telemetry methods) is adequate for some purposes, but not (because of limited bandwidth) for handling data from many of the emerging interdisciplinary instrumentation platforms, which collect several channels of data at high sampling rates. Other systems such as Inmarsat are prohibitively expensive for researchers.

A collaboratory facilitated by improved data communication capabilities and enabling the rapid exchange of in situ and satellite data at modest cost would help to meet one of the basic needs of oceanographic researchers. Ideally, such a system would have the capability to modify sampling rates in response to changing environmental conditions. For data from many coastal environments, transmission by cellular radio methods will suffice (e.g., Dickey et al., 1992). However, for data from the open ocean, satellite methods will most likely be needed.[2]

Cost-effective data telemetry from uncrewed platforms at sea would give oceanographers wider access to the ocean, the majority of which falls outside the commercial sea lanes. Two-way communication with remote platforms would permit conditional sampling and remote maintenance, further enhancing the value of the platforms.

Real-time access to all data as they are collected in the ocean has the potential to revolutionize field work. Data analysis and publication would no longer wait for the recovery of the instruments. Model testing and development could be carried out in parallel with the field work, perhaps guiding revisions to sampling strategies, rather than waiting for the release of data that usually postdates instrument recovery by one or more years. The instruments, if still working, would not need to be recovered, except, perhaps, for calibration or refurbishment. In some cases, it might be more economical to leave them in place rather than to field a recovery cruise. The design of the instruments themselves would also change, with on-board data storage hardware no longer needed or used only as a backup.

Providing Tools for Collaboration in Oceanography

The research tools of the oceanographic community—such as ships, satellites, moorings, drifters, supercomputers, and specialized modeling software—are often expensive, specialized, and inaccessible (often they are one of a kind). These resources could be put to better use if the tools and/or the results of their use were more widely accessible via electronic networks.

The needs of the oceanographic research community for supportive infrastructure can be described in terms of a number of tools that would better support collaboration and further improve access to data. The integration of those tools into a collaboratory could be particularly helpful to oceanographers.

TOGA Data Catalog

TOGA data sets are routinely distributed to the oceanographic and meteorological communities by magnetic tape, CD-ROM, electronic mail, dial-up databases, tabulations in monthly bulletins, and other media. Dissemination of these data has been encouraged by the U.S. and international TOGA project offices and represents the collective effort of individuals involved in field programs, specialized TOGA data centers, and national oceanographic data centers. Although current means of data distribution are adequate for many purposes, a centralized system for interactive, on-line access to information on essential TOGA data sets would greatly enhance collaborative work in short-term climate studies. Such a tool might be called a TOGA data catalog.

A TOGA data catalog could run on an X-windows workstation with a user-friendly point-and-click interface to allow interactive browsing of available data sets. Users would be linked by the Internet to the specialized TOGA data centers and to other providers of relevant data, so that the database could be updated regularly (every day to every month, depending on data type). Data would be classified by geographic location, depth range, time of collection, geophysical variable (e.g., ocean temperature), sensor platform (e.g., mooring, drifter), quality control and degree of processing (e.g., real-time vs. delayed mode), and by cruise (for shipboard data). Analyzed fields from numerical ocean climate models such as that used at the U.S. National Meteorological Center (NMC) would be accessible, as would be the analyses distributed in hard-copy form on the pages of NOAA's Climate Diagnostics Bulletins.

Information on data sets would be displayed graphically where appropriate (e.g., on a geographical mock-up of the globe, for information on position). Each data summary would be annotated with information on the data sources and would provide instructions on how to access the data via the Internet. Display capabilities built into the data catalog would allow visualization of the data sets on a user's terminal. Displays would be in the form of time series, vertical sections, horizontal fields, movies,

and so on. Users would be encouraged to forward comments on their experiences in working with particular data sets, and these comments would be incorporated into the annotated database.

A prototype version of such a data visualization and analysis system being developed at NOAA's Pacific Marine Environmental Laboratory is focused at the moment on data from the TOGA Tropical Atmosphere and Ocean (TAO) moored array and NMC operational model analyses. The TOGA TAO workstation concept is expandable in principal to encompass a much broader spectrum of data sets relevant to climate studies.

Globe Data Catalog

Data collected in oceanographic research are often difficult to find and/or to access. Several oceanographic data catalogs exist now, but each is based on its own paradigm. A globe data catalog—a system that could be accessed easily with a workstation featuring advanced three-dimensional graphics—would provide all oceanographers a common means to access various data types and their locations and periods of collection (Box 2.4). One goal of such a system would be to facilitate collaboration by simplifying access to data and providing for all oceanographers the kind of shared experience that fosters understanding and trust. The challenge is to design an earth sciences data catalog paradigm that all oceanographers would refer to when discussing their data with one another.

A system of the kind described in Box 2.4 avoids the problem of data format completely. Instead, the focus is on the definition of supported views. Supported views might include a standard form

BOX 2.4 GLOBE DATA CATALOG

Imagine a three-dimensional representation of the globe displayed on the scientific workstation sitting on your desk. (If you want to be imaginative, consider it as a hologram spinning around slowly in a large room!) Imagine further that any oceanographer in the world can see the same representation simply by typing a single command at his or her computer.

Spotted over the globe are a scattering of dots, icons, and square grids representing the locations where various types of oceanographic data have been collected. These "data catalog objects" can be organized into layers representing different types of data. Multiple layers can be shown at the same time by displaying each as a different color. Shown are some standard layers defined by oceanographers, ecologists, biologists, and others. Individuals can define their own layers if they wish.

Across the top of the workstation screen are the date and the time. A user can scroll forward and backward in time. Scrolling 100 years or so backward might reveal icons on the globe that identify where to find the ship logs kept for long-ago cruises. Scrolling forward in time to the future might display locations for future cruises. For the time between, a catalog of oceanographic information is represented.

Each data catalog object contains data descriptions of one or more "data catalog items" representing various data sets. The data set itself is not stored in the system. Instead, views of data are stored. A view is someone's interpretation of the data. The first view produced for a data item might simply be a reference to how the data set was collected, who collected it, and where it is now held. Another view might be a Tektronix 4010 plot of the data. A fancier view might be an X-windows movie visualizing a time series of the data. In some cases the algorithm used to generate a view might be included somehow as part of the view.

Data catalog objects are selected with a mouse either by pointing at a data catalog object and clicking on it or by somehow dragging a mouse through many dimensions—latitude, longitude, depth, time, and multiple type-layers. Selecting data objects on the globe generates a menu of the related available views of the data objects selected, along with information on who published a view. As mentioned above, layers might be defined by data type, discipline, author, instrument type, or some other category. Users of the system would provide their own profiles defining how the layers would look for each individual.

submitted to a data archive describing the data set, a Tektronix plot file, a subset of the data collected, or a message to connect to some person's computer on the network and run a program, which in turn would generate an X-windows session on a user's workstation.

Note that security is an important part of such a system. In fact, in some cases, attempts to access a view would produce the message "Access denied; please call Mr. Jones for further details." Groups of collaborators would be assigned access codes so that they could access each other's views. The details of how to provide reliable security for such a system would be an interesting area to pursue.

If any oceanographer on a network can bring up a globe data catalog on his/her screen, then the catalog resembles a publication. In fact, it provides an opportunity for an individual to publish his/her data and to get credit for the task of collecting the data. A view would thus be another type of authored publication with the appropriate amount of credit attached to it. Perhaps the system could employ hypertext capabilities such that the number of links back to another person's data-view would deserve credit as well.

Data set editors would be assigned to data catalog objects. Each would evaluate the views submitted for the data and assure the quality of each. In a sense this person would have the same role as a journal editor.

Although the technological aspects of such a tool are fun to consider, an important contribution, in terms of increasing collaboration, would be the development of a tool that provides a common experience for all oceanographers to work from.

Tour Tool

Many useful resources are already available on the Internet. The problem is knowing whether they are useful for doing one's own research. Needed is an easy-to-use tool that would allow researchers to generate a tour or "sample session" in the use of applications available on the Internet. Such a tour would be similar to those tours Claris Corp., a subsidiary of Apple Computer Inc., uses to introduce products to new users.

A tour tool would enable a user to access a tour available on a node on the network and then sit back and simply watch two or three sample sessions using the network resource of interest. Tours would require different levels of user equipment and software and would have to be advertised as such. Tours might include, for example, (1) a sample anonymous session on a Unix workstation, (2) accessing an oceanographic database service such as the Ocean Network Information Center (OCEANIC) maintained by the University of Delaware, (3) sample library searches on various library systems, and (4) an X-windows demonstration using the Lamont-Doherty View-server.

Those interested in creating a tour would need a session-recorder or movie maker (in fact, session recorders for several platforms) and perhaps a tour-file format to generate tours or sample sessions of the resources they provide on the Internet. Such a tool might provide the capabilities to easily annotate a tour with text, annotate with voice or music, record mouse movements where appropriate, or give users a chance to try doing something simple, interactively.

Cruise Planning Tool

Oceanographers often collaborate through joint participation in research cruises that bring together scientists from diverse disciplinary areas and from many countries. Logistical and scientific planning for cruises could be facilitated by a tool with a variety of functions, including (1) a capability for determining space availability and utility; (2) a calendar for ship scheduling; (3) sampling regimens; (4) forms, including shipping manifests, crew medical and personal information forms, and State Department forms for clearance to work in foreign waters; (5) sample documents showing a planned cruise track; and (6)

an agreement indicating the breakdown of ship time between different investigators on board. The concepts underlying such a cruise planning tool could be applied to tools for use with distributed oceanographic facilities, including ships, moorings, drifters, field laboratory facilities, and others.

Ontology Tool

As used during discussions at the CSTB workshop, the term "ontology" connotes a set of perceptions, terminology, procedures, and perhaps myths that are common to a group of people. Oceanographers within subdisciplines (e.g., physical oceanography) have particular ontologies, just as oceanographers in general have ontologies that may differ from those of their mother disciplines (e.g., physics). In order to collaborate on interdisciplinary problems, oceanographers need to develop an ability to communicate using a common set of terms. An ontology tool could be developed using network and computer resources to facilitate the introduction of an individual to the unfamiliar ontology of collaborating oceanographers. To be effective, such an ontology tool might have to include videos and face-to-face experiences that would convey visual effects and other subtleties.

ATTRIBUTES OF A USEFUL COLLABORATORY FOR OCEANOGRAPHY

A collaboratory for oceanographic research will require a variety of information resources and services that support a broad spectrum of interactions ranging from informal communication to active collaboration, possibly on interdisciplinary research, and including "passive" collaboration via controlled databases. Such a system would consist of a suite of tools whose usefulness will depend critically on the degree to which several criteria are satisfied: interoperability, transparency, customizability, integrity, and extensibility. Although these criteria are desirable even for the less formal modes of interaction, they are essential for the more formal ones.

Interoperability

It is imperative that the various components of oceanographic research—experiments, data, models, graphical and tabular interpretations, textual summaries, documentation, and publications—be thoroughly integrated and interoperable. This requirement is only partially satisfied by the imposition of standard data formats and interfaces. Portability of computational models, for example, is becoming an increasingly important issue as current parallel processing practice requires that programs be customized for each type, brand, and generation of machine. Portability, an active area of computer science research that is not simply a matter of standards, is highly relevant to the notion of a collaboratory if models are to be shared.

Data and the tools for their interpretation pose a number of research issues that also go beyond mere standards. Researchers may wish to view or process data in ways that the initial investigator did not imagine. The organization of scientific databases to allow flexible access to multidimensional, temporal data is also a research area that underpins the development of a collaboratory. This need is particularly acute in interdisciplinary work involving biological and chemical as well as physical data possibly collected over a range of time and space scales.

Effective and proper use and interpretation of data submitted to a collaboratory will require adequate documentation of the experiment in which they were collected. While these "metadata" may customarily be provided as informal annotations, ideally they, too, should be formalized to permit modeling and analysis software to assess possible inappropriate use of the associated data.

Transparency

Related to interoperability, the second major criterion of a useful collaboratory is transparency. The collaboratory will likely be implemented as a physically distributed system with individual institutions or researchers perhaps maintaining "ownership" of their own contributions and providing network access to others. Ideally, the distributed nature of the information will be transparent to users, thus obviating the need for explicit connections to remote machines and explicit transfers. Access time can be reduced through general mechanisms such as fast networks, caching, and replication. It has been argued by some in the oceanography community that there is no compelling need to provide on-line access to raw data, but rather only to summaries, analysis, and interpretation. The argument is that the data may always be obtained informally, if necessary, for further investigations, but that maintaining general access to all data would overwhelm the system. In a transparent system, however, the data could remain accessible without burdening the system or the users.

Customizability

A properly designed and successful collaboratory must also recognize the needs of its users, both as individuals and as identifiable subgroups. Users may wish to have customized interfaces that allow interaction on familiar terms. Multiple degrees of collaboration, including single users, enumerated groups, and classes of users, should also be supported to protect information not ready for full disclosure. User-imposed restrictions on the scope of information of interest could potentially be used by the system to optimize access in a distributed environment.

Integrity and Extensibility

Integrity refers not only to maintaining system and data security across the various partitioned layers of a system, but also to assuring the validity of information submitted to the system and thus of subsequent research based on that information. Although protection of integrity depends on policies and procedures at least as much as on technology, features such as formalization of metadata and documentation of models can provide valuable support for ensuring integrity.

Finally, a national collaboratory must be extensible to allow for the consistent incorporation of new services and features. Ensuring extensibility requires that considerable forethought be given to all representation and interface standards so as not to preclude or inhibit any foreseeable extension of the system.

NOTES

1. The lack of in situ data and difficulties associated with collecting more will not be resolved by the exclusive use of space-borne oceanographic sensors. Satellite sensors can indicate sea surface temperature, wind direction and velocity at the surface of the ocean, ocean color (indicative of chlorophyll content), the height of the sea surface (which, in part, reflects the currents in the ocean), and other oceanographic variables, but they cannot see into the ocean. Moreover, although satellites often give a more densely sampled and more global data set than do ships and buoys, remote sensing methods rely critically on calibration by the best available in situ data, collected in a timely fashion by ships and buoys at many locations, to achieve accuracy and reliability. For example, unless they are confirmed by ground truth data, sea surface temperatures sensed by Advanced Very High Resolution Radiometers (AVHRRs) on NOAA polar-orbiting satellites may be in error in excess of 1°C due to contamination of the atmosphere by erupting volcanoes such as Mt. Pinatubo. In addition, AVHRR and satellite color sensors can provide no useful data during cloudy conditions or at depths below a few meters.

2. Note that Brooks and Briscoe (1991) have reported success with high-frequency radio transmission also. A possible solution to the satellite data transmission problem is the use of excess capacity on a communications satellite that is part of NASA's Tracking and Data Relay Satellite System (TDRSS). Data transmission costs might be reduced if the excess capacity of a NASA satellite could be harnessed for oceanography, but many issues remained unanswered as to the feasibility of using TDRSS.

3

Building Collaboratories in Space Physics

Space physicists study the interaction of charged particles with electric and magnetic fields in the space environment. The interactions are complex because the motions and distributions of the charged particles are determined by the fields, but at the same time, those distributions and motions affect or determine the electric and magnetic fields. Space physics research has many aspects: (1) fundamental physics, in which the goal of the research is to understand the types of interactions that occur; (2) a phenomenological component aimed at describing and understanding the distribution and behavior of plasmas in settings such as Earth's magnetosphere; and (3) attention to applications, for example, attempts to predict the occurrence and nature of disturbances in Earth's plasma environment in order to prevent interference with or damage to systems such as power grids or Earth-orbiting satellites.

With advances in technology has come an increase in the magnetosphere's influence on human activities. Our communications, weather, surveillance, and test ban verification satellites are stationed deep in an active region of the magnetosphere. Communications at many wavelengths are affected by ionospheric conditions, which in turn are affected by magnetospheric conditions. Every solar cycle since the 1930s has brought electrical power system disruptions caused by disturbances in the magnetosphere, a relationship acknowledged only over the most recent solar cycles. Pipeline corrosion is accelerated by currents induced by rapid magnetic changes. Magnetic prospecting and other activities may be helped or hindered, depending on magnetospheric activity at the time.

Understanding the cause and the nature of magnetospheric disturbances is thus an important endeavor supported by the National Science Foundation and the National Aeronautics and Space Administration, as well as the National Oceanic and Atmospheric Administration, Department of Defense, and Department of Energy, and requiring both ground- and space-based observations of electric and magnetic fields, waves, plasmas, and energetic particles. Interpreting the information gathered generally involves joint collaborative analysis of two or more different data sets. Thus collaborative studies are frequent in space physics, if not generally imperative.

This chapter briefly outlines the types of data collected and the instrumentation, methodology, and techniques of analysis used in the field of space physics; reviews some of the initial collaborative efforts in space physics research and describes additional ongoing collaborative programs; and suggests how a national collaboratory for space physics might benefit researchers and, in turn, what components scientists would regard as basic to a useful collaboratory.

SPACE PHYSICS RESEARCH

Data Collection and Instrumentation

The earliest recorded observations of phenomena now studied by space physicists were sightings of the aurora. Later, when the compass became a precise tool for navigation, short-period (seconds to hours) fluctuations in Earth's magnetic field were recorded and the existence of an ionosphere was

inferred. Early in the 1900s, over-the-horizon radio transmissions proved the ionosphere's existence. The alignment of auroral forms with Earth's magnetic field and the correlation of auroral activity with solar activity deduced in the late 19th century were the first real indications of the existence of a magnetosphere and its control, in some unknown manner, by the Sun. In 1930 Sidney Chapman and A.C. Ferraro developed the first real model of the magnetosphere. They postulated that a plasma (which we now called the solar wind) was emitted intermittently to cause a compression of Earth's magnetic field and a resultant magnetic cavity in the flowing solar wind (which we now call the magnetosphere). By the mid-1950s the plasma density of the magnetosphere had been probed remotely using lightning-generated, very-low-frequency waves. Nevertheless, the beginning of space physics as a discipline was marked by the discovery of the Van Allen radiation belt and the in situ exploration of Earth's magnetosphere and its interaction with the solar wind made possible with the advent of rockets and satellites in the late 1950s and 1960s.

The data gathered in space-based observations are collected by sets of 3 to perhaps 10 different sensors positioned on Earth-orbiting satellites or interplanetary probes. A typical satellite might carry a magnetometer to detect slowly changing magnetic fields and a separate sensor to measure magnetic oscillations; a device to record electric fields and waves; plasma analyzers (often a coordinated set of sensors) to measure the fluxes of charged particles as functions of their mass, energy, and direction of motion; and one or more sensors to measure high-energy charged particles. The data consist of time sequences of the sensors' outputs. When possible, similar or complementary data are collected by other satellites in different locations in space.

In addition, data gathered simultaneously from ground-based instruments are used to examine phenomena such as disturbances in the geomagnetic field (as detected by arrays of ground-based magnetometers), changes in the ionospheric density (as indicated by radar, riometer, and rocket-based observations), enhancements in atmospheric emissions signifying excitation by energetic particles (as shown in images from photometers and all-sky cameras), and activity on the Sun (as shown by, e.g., coronagraphs). Data on particles and fields are now often augmented with images of Earth's atmosphere in the visible, ultraviolet, or x-ray regions of the electromagnetic spectrum, thus enabling space physicists to relate observed high-altitude phenomena to ground-based observations.

The data matrices produced by space- and ground-based instruments thus include many different kinds of measurements often taken at widely separated sites, but often with good (but differing) time resolution. A challenge in analyzing and interpreting the data is to combine and compare them so as to deduce a global picture of the behavior of the magnetospheric system. Combining and analyzing these various data sets and types both require extensive electronic communications, including electronic mail and network transfer of data and text files, between the home institutions (sometimes international) of the many investigators involved.

Methods and Technologies for Data Analysis

Space plasma physicists use a combination of analytical and numerical methods to interpret and understand data, as well as to assist in the planning of observational campaigns. These methods can range from the use of simple linearized equations to model the early time evolution of plasma waves to the use of extensive multidimensional codes that attempt to simulate the full complexity of these nonlinear systems. The simulations can consist of calculations of the locations and motions of many millions of individual charged particles, together with the electric and magnetic fields that they generate, or they can be solutions of simultaneous partial differential equations that describe the plasma as a "fluid." Running a typical simulation requires the use of a supercomputer, although some of the newer workstations can now run some of the smaller codes. The amount of numerical "data" generated is so great that if one were to attempt to keep it all, it would far exceed the capacity of even the largest computers. Even after pruning, the files that are kept necessitate good (reliable and fast) network communications between the

supercomputer and the researcher's home site, so that the "data" can be transferred between the two locations for detailed analysis, manipulation, or graphical display.

Simulations are now becoming a common part of space physics projects because the models are becoming sufficiently realistic to warrant direct comparisons with the measurements. Indeed, one of the greatest challenges in a number of current programs is the synthesis of models and measurements through coordinated displays and analyses. Thus, manipulation and display of diverse data sets from both numerical modeling and a wide range of experiments are a focus for computational and telecommunications activities in space physics.

EXAMPLES OF COLLABORATIVE EFFORTS IN SPACE PHYSICS RESEARCH

Of the initial collaborative efforts in space physics research, four are examined below: (1) the Space Physics Analysis Network, developed to satisfy the need of researchers to be in close and rapid electronic contact with collaborating scientists; (2) Coordinated Data Analysis Workshops, a response to the need to synthesize extensive data sets into succinct scientific results; (3) the Active Magnetospheric Particle Tracer Explorer/Charge Composition Explorer, a scientific mission with a strong, centralized facility for data analysis; and (4) the Sondre Stromfjord Observatory testbed, a new effort that, when complete, will allow scientists to operate their instruments in Greenland remotely from sites in the United States. Also discussed are the Solar-Terrestrial Energy Program, NSF's Geospace Environment Modeling Program, and the International Solar-Terrestrial Physics program, which provide additional evidence of the increasingly important role of collaboration in the space physics community.

The Space Physics Data System, envisioned as an aid for individual researchers and a first step toward building a broad-based system for handling space physics data, is outlined as an approach whose potential utility has been acknowledged within the community of researchers.

Initial and Ongoing Collaborative Programs

Space Physics Analysis Network

In September 1980 members of the space plasma physics science community from more than three dozen institutions met to discuss what steps were needed to provide a more coordinated approach to solving many data access and analysis problems. There was then, and still remains today, a strong need to better utilize existing and diverse space physics databases through collaborations with remotely distributed space scientists. This initial working group recognized that the technology existed to interconnect distributed computer systems that would provide the "enabling environment" for remotely accessing databases at a low cost (Greenstadt and Green, 1981). It was quickly realized that a computer-to-computer communication system could be achieved (which satisfied many of the desired objectives even though significant funding was not available) by making maximum use of existing computers, equipment, and facilities at the remotely distributed sites. Within a year after that initial meeting, the Space Physics Analysis Network (SPAN) became operational with three nodes at locations in Alabama, Texas, and Utah. Almost immediately, many scientists became involved in previously impossible collaborative activities such as simply comparing data on the same time scale (Green et al., 1983; Rees et al., 1986; Sanderson, 1990; Thomas and Green, 1988a,b; and Thomas et al., 1987).

Within 8 years, SPAN grew rapidly to include several thousand computer nodes in the United States and was extended to Japan and many countries in Europe and South America. The network relied explicitly and to an unprecedented degree on the active involvement of its users.

A scientific oversight group, the Data Systems Users Working Group, guided the entire effort, meeting approximately every 9 months (see, for example, Baker et al., 1984). The actual operation of

the network required extensive volunteer labor from the institutions on the network, which were also represented in the users working group. This approach linked the users and developers in a tightly coupled feedback loop and enabled the network to meet many of the users' highest-priority needs, greatly enhancing many NASA and non-NASA space science programs (Green and King, 1986; Thomas and Green, 1987; and Winterhalter, 1986).

SPAN succeeded far beyond the dreams of its initial developers. Its success was recognized both by participants and by those responsible at NASA for electronic communication. Eventually, the SPAN effort was given a more permanent home as part of NASA's contribution to and participation in the Internet. The new NASA Science Internet currently preserves the original SPAN functionality and has added TCP/IP connectivity.

Coordinated Data Analysis Workshops

The concept of the Coordinated Data Analysis Workshop (CDAW) arose following the highly successful data-gathering phase of the International Magnetospheric Study from 1976 to 1978. The CDAW's purpose was to bring together all available magnetospheric data for specific periods of time to determine how a particular magnetospheric process worked. Nine CDAWs have been held to date. Recent workshops have focused on global-scale solar-terrestrial physics problems requiring diverse data sets and a variety of modeling skills. CDAWs in general have involved a broad cross-section of the space physics community (especially the solar wind-magnetosphere-ionosphere community). The initial phase of a CDAW effort is selection of the problem to be addressed; relevant data are then identified and collected from multiple (as many as 10) spacecraft and from scores of ground-based facilities. These data are then sent to the National Space Science Data Center (NSSDC) at Goddard Space Flight Center to be placed into a common data format. The resulting databases have often included literally hundreds of individual data sets for selected analysis intervals. Those who contributed data are then invited to gather at a common site and jointly analyze aspects of the problem at hand.

Early CDAWs were "paper" workshops—participants simply brought data records plotted on a common time scale for comparison with other data. The more recent workshops have utilized a common database accessed by interactive computer systems used during the workshops, held normally at the NSSDC but also at other locations such as Stanford University and Toyokawa, Japan. The face-to-face workshops, themselves a key element of the overall CDAW process, have had as few as a dozen participants (in splinter CDAW meetings) or as many as 100 participants in the major CDAW meetings. After the workshops, the participants return to their home institutions to prepare joint oral and written papers on their results. Recently this process has been aided by the dissemination of the CDAW databases on CD-ROM.

A major aspect of CDAWs has been the cooperative sharing of expertise. Such sharing has been facilitated by the face-to-face workshop format, which allows scientists who sometimes have very different views of the physics involved to discuss alternative interpretations of the same data sets. Perhaps an even more significant aspect of CDAWs has been the open sharing of data (prior to publication) obtained from many different instruments. The sharing aspect of CDAWs has been greatly aided by the adoption of "rules of the road" as given in Appendix C.

One of the major difficulties associated with CDAWs is the large overhead cost involved in assembling the infrastructure needed for the collaboration, including the assembling of the data and the coordinating of meetings and people. Thus far fewer workshops have been held than the space physics community would find most beneficial.

Active Magnetospheric Particle Tracer Explorer/Charge Composition Explorer

In the fall of 1984, the Active Magnetospheric Particle Tracer Explorer/Charge Composition Explorer (AMPTE/CCE) spacecraft was launched into Earth orbit in the fall of 1984 as part of a three-spacecraft Explorer mission to study plasma interactions in naturally occurring plasmas and in those created in space through the release of chemicals. It carried five scientific instruments that made measurements of the electric and magnetic fields and plasmas in Earth's magnetosphere. The CCE was designed to measure the ions of barium and lithium that entered the magnetosphere from releases in the solar wind and the magnetotail. This unique experiment indicated that barium was being released from a large portion of Earth's surface.

The complementarity of the spacecraft experiments (the instruments each gave quite distinct measurements), the relatively small size of the science team (~20 to 30 scientists), and the management of the science mission from a single institution provided an important opportunity to develop a centrally located but remotely accessed data analysis system that fostered scientific collaboration. This central facility produced raw data from spacecraft telemetry, provided these data and information about spacecraft position to all remote institutions, and processed and distributed survey data products. Remote institutions accessed the data through dedicated telephone lines and SPAN, allowing individual scientists to further process data both at the central facility and at their home institutions. Thus, those with access to the database could work with others without incurring the overhead costs involved in contacting many institutions.

This mission was successful because it provided almost effortless access to a reasonably well funded, well-managed central data facility, as well as access to a relatively small number of dedicated scientists who understood individual instrument performance and were interested in scientific collaboration, and the freedom to analyze data remotely using tools available at home institutions rather than depend solely on the standard analysis routines provided by the central facility.[1]

Sondre Stromfjord Observatory Testbed

Several existing ground-based observatories, such as the Sondre Stromfjord Observatory in Greenland and the numerous stations in Antarctica used to study upper-atmosphere and space physics, offer only limited access because they are located in remote regions of the world. The space physics community and other researchers would benefit enormously if these instruments could be operated remotely via computer networks, thus improving access to real-time data.

The Sondre Stromfjord Observatory in Greenland is currently being used as a testbed for enabling collaborative research. The facility has many attributes that make it an excellent choice for a collaboratory project: it is remote, the user community is distributed and manageable in size, and it already has a modest networking infrastructure in place.

The real-time viewing of data obtained at the Sondre Stromfjord Observatory will allow scientists to perform experiments that previously required on-site decision making, allowing those not present at the remote observatory to respond, for example, to rarely occurring geophysical phenomena such as solar proton events, which affect ozone depletion and represent a hazard to aircraft flying on transpolar routes. Also, the use of computer networks will enable more experimenters to be involved simultaneously in research at the remote facility, thus enhancing decision making and productivity. The possibility of remotely adapting preplanned experiments to existing geophysical conditions is also an exciting prospect.

Solar-Terrestrial Energy Program

The worldwide community of solar-terrestrial scientists has embarked on an exciting and intellectually rewarding project: to understand quantitatively the linkages from the Sun through the interplanetary medium and into the depths of the surrounding geospace. The variety and complexity of the physical processes involved in these linkages have challenged our ability to understand the total system. Now, through a concerted global effort, the Solar-Terrestrial Energy Program (STEP) has begun to use remarkable new observational tools and modeling capabilities to achieve an unprecedented comprehension of our solar-terrestrial system. STEP was approved by the Scientific Committee of Solar-Terrestrial Physics in 1986 and launched in 1990; the International Council of Scientific Unions gave its formal endorsement in 1987 and recently extended the program through 1997.

The main scientific goal of STEP is to advance the quantitative understanding of the coupling mechanisms responsible for the transfer of energy and mass from one region of the solar-terrestrial system to another; the main practical goal is to improve the ability to predict the effects of the variable components of solar energy and mass flows on the terrestrial environment, on technological systems in space and on Earth, and on the biosphere.

A well-coordinated ground- and space-based observing program is essential to accomplish these goals. Basic in situ measurements will be obtained by the various spacecraft missions approved by the Inter-Agency Consultative Group as a cooperative project of the world's four major space agencies. In parallel with these efforts, STEP will coordinate the use of specially designed ground-based instruments and aircraft, balloon, and rocket experiments and will promote theory development, modeling, and simulation studies on an international scale.

Crucial to the success of STEP are the dedicated information and data systems that allow scientists from all participating countries to improve communications among themselves and facilitate the coordination and standardization of measurements, as well as the interchange and analysis of data. These systems include those operating at satellite situation centers and coordinated data-handling facilities, at new ground-based observation situation centers and computer simulation centers, and during coordinated data analysis workshops. These facilities should provide sufficient support at present for the active programs, but the problem of accessing and studying data from older missions remains. Toward the end of the STEP mission, when the data acquisition phase is complete, this problem will apply also to currently active programs.

Geospace Environment Modeling Program

Geospace encompasses the regions from Earth's upper atmosphere to the Sun. The Geospace Environment Modeling (GEM) program at NSF is an effort to study the near-Earth portion of geospace ranging from the lower ionosphere to the environment where Earth interacts with the solar wind. This space plasma environment, called the magnetosphere, is an electrodynamic system that links Earth's atmosphere with the local astrophysical system; thus magnetospheric physics as a discipline lies at the intersection of Earth-system science and modern astronomy.

The purpose of GEM is to enable research, especially collaborative research, on the dynamical and structural properties of geospace that will lead to the construction of a geospace general circulation model with predictive capability. GEM pursues this goal by supporting ground- and space-based observational research, as well as theoretical research, in modeling of the geospace environment. A geospace general circulation model is analogous to general circulation models of the lower atmosphere and will require extensive interdisciplinary and intradisciplinary research.

The strategy for achieving the GEM goal is to undertake a series of campaigns involving theoretical and observational research focusing on particular aspects of the geospace environment such as the magnetotail and substorms, global plasma models, cusp signatures, and electrodynamic coupling.

The first phase of GEM, which is now in progress, focuses on the magnetospheric cusp and boundary layers. The next phase of the program, scheduled to begin in 1994, focuses on the magnetotail and substorms. For each campaign, working groups of scientists are formed to analyze the data and develop the computer-based visualization and simulation programs needed to study given phenomena. These programs will be developed in a modular fashion with the intention of building a flexible yet robust research model of near-Earth geospace—the most highly coupled of geospheres—that will be available to other members of the scientific community.

International Solar-Terrestrial Physics Program

The International Solar-Terrestrial Physics (ISTP) program is the major research effort in space plasma physics for the decade of the 1990s. It began in the early 1980s as a NASA four-spacecraft mission called Origins of Plasma in Earth's Neighborhood (OPEN); the goal was to monitor the flow of mass, momentum, and energy from the solar wind through the magnetosphere into the upper terrestrial atmosphere. Since that original OPEN concept, the program has evolved in many ways. Japan took over responsibility for one of the spacecraft (GEOTAIL), and the European Space Agency has put two of its satellite programs (SOHO and CLUSTER) into the overall ISTP mission set. Moreover, Russia is also planning to coordinate some of the former-Soviet Union's independently planned satellite missions, such as INTERBALL, with ISTP, thus leading to well over a dozen highly instrumented spacecraft that will contribute to ISTP. This international collaboration is being coordinated by the four major space agencies through the Inter-Agency Consultative Group.

ISTP is unique in the degree to which ground-based measurements and theory will play a central role. Furthermore, the observations most critical for collaborative studies, whether collected by space- or ground-based instruments, will be placed into a "key parameter" database that will be accessible almost instantaneously to all participating scientists, who will share a huge, common database of high-quality measurements, analysis and display tools, and special models. The ISTP program will thus provide what will be tantamount to a continuing Coordinated Data Analysis Workshop environment for analysis of space physics data.

ISTP will continue throughout the 1990s as new spacecraft are launched. However, in the period from 1993 to 1996 there will be unique opportunities to assemble global geospace data sets in conjunction with particularly intense data collection campaigns and large-scale analysis efforts. The various world space agencies are planning now, through a series of workshops, the optimum ways to acquire, analyze, and disseminate the ISTP data. It is expected that the data systems now being set in place will suffice to facilitate the required collaborative studies during the active phase of the ISTP program. However, it is not certain how the infrastructure support needed in the post-project period will be provided.

Space Physics Data System: A Collaboratory of One

Novel ideas for and approaches to scientific investigation in space physics, as in most disciplines, generally start with the efforts of a single, highly motivated individual, who then, when appropriate, enlists and encourages the participation of his/her peers in the investigation. In space physics, this participation—at both the individual- and multiple-investigator level—generally takes the form of amassing and analyzing data from many different instruments carried on one or more spacecraft.

Certain boundary conditions are thus demanded of a system that supports space physics data analysis. The first set of requirements exists at the individual-scientist level and can be called a "collaboratory of one." Before individuals can participate in a larger collaboratory effort, they must first have their own well-developed ideas that they can communicate to their peers. Thus individual researchers must have ready access to multiple data sources and computational platforms whose use

should not require labor-intensive programming, extensive and arcane methods for locating and transferring data, or detailed knowledge of how the data were initially acquired. An ideal collaboratory of one, the so-called Space Physics Data System, should allow a space physicist to:

- Search a global mega-database to locate all data available;
- Transparently transfer the actual data without knowing the details of the transport methodology;
- Convert the data to scientific units, without having to program, by using a set of generic software tools tied to a standardized, self-documenting file system;
- Display and manipulate the multiple data sets via a set of standardized, X-windows-based tools either locally or via network; and
- Easily generate value-added analysis and display software tools in a standardized paradigm that allows further use of the derived data sets.

A collaboratory of one that enables such actions would allow individual investigators to use their limited resources for actual scientific analysis as opposed to dealing with "computerese."

Such a Space Physics Data System would facilitate the next step to a joint, multiperson collaboratory requiring a system to accommodate the sharing of derived products and ideas. It is well within the capabilities of current workstation and network technology. In the future, as multimedia technology becomes widespread, sharing of a richer form of ideas and analysis will become possible through the sharing of graphics, text, video, and sound. Thus the collaboratory of one is a necessary first step in the process of developing a global Space Physics Data System involving multiple sources of data and their use at geographically distributed sites. A movement toward such a system has begun within the community with active encouragement from NASA, but to date these efforts have not reached even the pilot project stage.

NEW OPPORTUNITIES PROVIDED BY A COLLABORATORY FOR SPACE PHYSICS

A collaboratory for space physics would provide researchers with a number of opportunities to make new and better use of data acquired by space-based, suborbital, and ground-based experiments. Currently funded investigators could use on-line directories, user-friendly access interfaces, and network file transfer to obtain correlative data on the state of the Sun, solar wind, or ground geomagnetic indices more simply than can be done today. Associated data from various instruments on the same spacecraft could be obtained through a project data request and dissemination system interfacing with the collaboratory network.

A more important use of a collaboratory would be for group studies of global problems. By definition, global studies require data from many instruments and many locations. Such data are invariably of many different types and formats and are difficult for any one individual to assemble and display. A collaboratory could provide access to facilities, support staff, computer hardware, and specialized software to organize and document data so that they could be quickly accessed and displayed by study participants. Case history studies of complex phenomena such as magnetic storms or magnetospheric substorms are an example of potential group efforts. The assimilation of multiple data sets into self-consistent maps of various parameters is another example. In magnetospheric physics the Assimilative Mapping of Ionospheric Electrodynamics model does mapping using the equations that govern ionospheric currents, field-aligned currents, electrical conductivity, and electric fields.

A collaboratory could also be used as a tool in planning multi-instrument, multi-platform campaigns. Specialized tools such as geographic information systems, software for predicting when spacecraft will be at key observing locations in the magnetosphere, and empirical magnetic field models could be brought together in a common computer system. These could be combined with the best available empirical or theoretical models of a phenomenon to predict when and where certain types of observations could be made. The use of particular instruments could be planned, telemetry tracking organized, and specialized ground measurements scheduled.

A similar application would be the capability to provide for a quick response to unexpected or transient phenomena. For example, the development of a particularly active region on the Sun often leads to very large magnetic storms that can severely damage or even destroy spacecraft and ground systems. A collaboratory might be used by a group of experts to quickly bring together data that could be used to predict the occurrence of such activity and perhaps anticipate possible effects.

A collaboratory might also serve as a repository and distribution system for software. Many data formats, data access routines, data analysis systems, and specialized transformation utilities that are developed at public expense are potentially of great value to others. A collaboratory could provide a directory to such software as well as a repository for it and its accompanying documentation.

It is conceivable that a collaboratory could be used for electronic publishing, or at least as an initial step in this direction. Today a central location such as the National Space Science Data Center could be used to receive formatted documents and graphics. Standards would have to be defined and a means provided for translating among the various common formats. Initially an electronic publishing center might be used only for documentation of the data for group studies and software descriptions. Eventually it might serve as a testbed for a broader vision of electronic publishing and literature distribution.

A collaboratory in space physics could be of use to scientists at small universities and colleges who might, for example, use a bulletin board to advertise their interest in cooperating on a research project of a given type. Interested researchers at a funded institution could respond via the bulletin board. Eventually the participants might exchange data through the collaboratory and use the hardware, software, and tools of the collaboratory to analyze the data.

Another possibility is that after the primary research has been done by the initial investigator, the databases developed for specialized studies could be written to CD-ROMs along with appropriate software to access and display the data. Such collections, which might also include electronic versions of the papers produced in the studies, could be used to carry out additional research. Alternatively, teachers or professors at various institutions could make use of the data and accompanying papers to develop problems and exercises for classes in science and mathematics. Yet another possibility is to use a collaboratory to provide a source of general information in science. A collaboratory might have a public bulletin board or user interface that could provide access to general information about space science.

COMPONENTS OF A COLLABORATORY INFRASTRUCTURE

The possible uses of a collaboratory for space physics suggest a number of components that would serve to enable efficient collaboration between two or more space physicists and even "collaborations of one" allowing a single investigator to synthesize data from a number of different sources. Above all, a collaboratory infrastructure for space physics must allow for easy and rapid access to a multitude of different types of data from numerous instruments on a number of different space- and ground-based platforms. In addition,

1. The system must provide for the education of its users and minimize the learning curve required to become proficient; easily accessible help services (including a telephone number) should be provided.

2. Standards are a cornerstone of a national collaboratory infrastructure. Even though it is not necessary that all data be in identical format, it is necessary that the data be translatable into a compatible format for seamless integration into data-handling and analysis systems. Standards for user interfaces, network protocols, and so on are fundamental to a collaboratory.

3. It is important that users of a collaboratory adopt "rules of the road" to protect the originators of data and ideas and to assure that collaborations are carried out in fairness to all participants and contributors, and that data, analysis tools, and ideas are used correctly.

4. Given the dependence of the space physics community on space-borne instrumentation and remote land-based observatories and facilities, it is essential that agreed-upon methods be developed for the remote operation of these instruments and facilities, especially in support of rapid-response campaigns of coordinated observations.

5. Networking is the backbone of the collaboratory infrastructure, because it provides the digital communications required for the sharing of data and ideas. Networking services needed include electronic mail, file transfer, remote log-on and execution, database management, teleconferencing, "whiteboard" or electronic "scratchpad" capability, shared access to common graphical displays, and so on. The networking services must all meet basic levels of performance with respect to quality, reliability, and response time, and higher levels of performance should be worked toward.

6. Electronic mail, a fundamental networking service required by a collaboratory, deserves special mention. Needed is the capability to transfer compound documents, that is, those with special characters and formatting as well as graphics, in a standardized form that is compatible with common word processing and desktop publishing tools. Support services in the form of directories, yellow pages, and so on are also needed.

7. Computer-supported cooperative work tools are important for access to shared resources such as data and computing resources.

8. Common access to software libraries is an important component of a national collaboratory. It is especially important that shared software be reliable, portable, and unambiguous as to its use.

9. The collaboratory infrastructure and its analysis tools should be directly available for supporting the education of space physicists and, more generally, students in the physical sciences.

10. The infrastructure must be affordable so that (a) it will be built and (b) it will not displace funding otherwise earmarked for space physics research. The collaboratory infrastructure must be regarded by the space physics community as a set of services well worth their support and as a fundamental part of their suite of research tools, or else it will be considered a funding burden.

NOTE

1. Despite its success at the time, the AMPTE mission is no longer supported as an active NASA effort, although the data sets, equipment, and software to support collaborative studies still exist.

4

Building Collaboratories for Molecular Biology

This chapter discusses how to facilitate some aspects of molecular biology research from the viewpoint of a computational biologist or an information scientist interested in constructing biological information systems. Electronic collaboration in molecular biology is facilitated currently by a variety of shared database systems. In addition several prototypes that support collaboration are in use now, and a model collaboratory, the Worm Community System, illustrates a potential future direction for community information systems that support retrieval, analysis, and sharing of essential data and literature.

Molecular biology is the analysis and description of biological organisms at a molecular level. To attempt to relate the observed function to the underlying structure, an immense amount of data must be considered, at many different levels, from studies of many different organisms. Molecular biologists have made many important theoretical advances from such comparisons, notably elucidation of the central role of DNA (deoxyribonucleic acid) as the carrier of genetic information in all free-living organisms.

A DNA molecule provides the "blueprint for life." It is composed of combinations of 3 billion pairs of four chemicals known as nucleotides: adenine and thymine, or guanine and cytosine (Box 4.1). To simplify, the DNA in humans is divided up into 24 different chromosomes. About 10 percent is aggregated into genes, which code for proteins. The other approximately 90 percent has no known function. The complete amount of DNA that defines any living organism is known as its genome.

Molecular biology has traditionally been conducted on a small scale. Most laboratories have been small and have functioned independently, employing a handful of people and producing a moderate amount of data, often on a subject exclusive to each laboratory. Further, many experiments have been easily reproducible, requiring only modest equipment. In contrast to other scientific domains, such as space physics, direct sharing of results from experiments has not been critical, since the experiments could be simply and cheaply rerun. The level of detail in the published literature has been considered sufficient for the sharing of knowledge and for collaboration between laboratories. However, large-scale nucleotide sequencing directed at the human genome (3 billion nucleotides!) may call for centralized, highly automated facilities and new modes of deciding on the division of labor and of disseminating data.

From both a theoretical and practical perspective, molecular biology has been one of the great success stories of modern science. More than 60 percent of all National Institutes of Health grants—covering a wide range of problems in medicine and biology—propose use of the molecular methods of cloning, mapping, or sequencing to investigate problems. More than 70 percent of all articles indexed by the National Library of Medicine contain molecular biology subject headings.

At this time, the field of molecular biology is diverging. One branch addresses a range of integrated biological systems, concentrating on the function of protein molecules in the processes of DNA replication and mutagenesis, transcription, and translation.[1] This branch entails a high quota of individual imagination in theory building. The other branch, typified in Walter Gilbert's description of an emerging paradigm shift in biology (Box 4.2), concentrates on the role of DNA in the cell. It is this side of molecular biology that is data intensive and is embodied in the Human Genome Project, a burgeoning scientific endeavor producing vast amounts of new information that researchers must select from to advance their investigations.

BOX 4.1 FUNDAMENTALS OF GENOME RESEARCH

"Molecular biology is the discipline that demonstrated the relationship between genes and proteins. Molecular biologists determined that the gene is made of DNA (deoxyribonucleic acid)—that is, DNA is the hereditary material of all species. What is more, in what is now scientific legend, Crick and Watson determined in 1953 that the structure is a double helix and concluded correctly that this specific form is fundamental to DNA's function as the agent of storage and transfer of genetic information. In fact, in biology generally, shape determines properties—that is, structure almost always determines function.

"The DNA double helix is both elegant and simple. Each strand of the DNA double helix is a polymer consisting of four elements called nucleotides: A, T, C, and G (the abbreviations for adenine, thymine, cytosine, and guanine). The two strands of DNA are perfectly complementary: whenever there is a T on one strand, there is an A on the corresponding position on the other strand; whenever there is a G on one strand, there is a C on the corresponding position on the other. That is, T pairs with A, and G pairs with C. This complete redundancy accounts for how a cell can pass on a complete set of genetic information to each of its two daughter cells during cell division: the DNA double helix unravels, and each strand serves as a completely sufficient template upon which a second strand can be synthesized. In addition to providing an easy mechanism for the replication of DNA, the redundancy also provides great resiliency against loss or damage of information during the life of a cell. Such loss or damage of information, when it occurs, is the basis for biological mutations.

"From a computer scientist's point of view, the DNA double helix is a clever and robust information storage and transmission system. As computer scientists accustomed to dealing with a binary alphabet will immediately recognize, the four-letter alphabet of DNA is sufficient for encoding messages of arbitrary complexity.

"In brief, particular stretches of the DNA are copied directly into an intermediate molecule called RNA (ribonucleic acid, also composed of A, T, C, and G). RNA is then translated into a protein—which is again a linear chain, but one assembled from 20 different building blocks called amino acids. Each consecutive triplet of DNA elements specifies one amino acid in the protein chain. In this fashion, biology "reads" DNA (actually, the RNA copy of the DNA) . . . [as if it were a computer program].

"Once synthesized, the protein chain folds according to laws of physics into a specialized form, based on the particular properties and order of the amino acids (some of which are hydrophobic, some hydrophilic, some positively charged, and some negatively charged). Although this basic coding scheme is well understood, biologists are not yet able to predict accurately the shape in which the protein will fold.

"In total, the human genome (the totality of genetic information in each of us) contains about 3 billion nucleotides. These are distributed among 23 separate strands called chromosomes, each containing about 50 million to 250 million nucleotides. Each chromosome encodes about 10,000 to 50,000 genes.

"With the extraordinary advances in molecular biology over the past 20 years, it is now possible to read the specific sequences of individual genes and to predict (by means of the genetic code) the sequence of the proteins that they encode. A major challenge for molecular biology in the next decade will be to use this information to predict the actual biological function of these proteins."

SOURCE: Reprinted, with permission, from Lander et al. (1991), p. 35.

BOX 4.2 TOWARD A PARADIGM SHIFT IN BIOLOGY

"The new paradigm, now emerging, is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. . . .

"To use this flood of knowledge [that is resulting from the mapping and sequencing of human genes and the genes of model organisms], which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life.

"The next tenfold increase [5 years from now] in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need. . . .

"We must hook our individual computers into the worldwide network that gives us access to daily changes in the database and also makes immediate our communications with each other. The programs that display and analyze the material for us must be improved—and we must learn how to use them more effectively. Like the purchased kits [of molecular biological reagents], they will make our life easier, but also like the kits, we must understand enough of how they work to use them effectively."

SOURCE: Reprinted, with permission, from Gilbert (1991), p. 99.

GENOME RESEARCH IN MOLECULAR BIOLOGY

Within molecular biology, the importance of sharing data and the scale of research have increased for activities revolving around gene mapping and sequencing. The widespread use of automated DNA sequencing technology[2] is now permitting the nucleotide sequence for many genes and chromosome regions to be determined. This enormous amount of information has brought centralized database archives to the fore as major points of collaboration. The main data produced in genome projects are the sequences of DNA base pairs embodying patterns of great evolutionary, physiological, and medical interest. The discovery of meaningful patterns is complicated: first, the data available to support a particular line of research are often sparse, so that many different organisms (e.g., bacteria, yeast, worms, mice, and humans) must be studied, and second, the available theory of the organization of DNA sequences remains scant, so that hands-on experimentation by scientists is required.

As experimental results, DNA sequences themselves are useful to share directly, since similarities in sequence often imply similarities in function. However, the printed literature is no longer adequate for sharing such data, partly because of the economics of charges for the journal pages needed to print the long sequences, but largely because computer examination of the sequences is far more effective than human examination of the data on a printed page. As a result, it has become standard procedure for investigators to use sequence databases such as GenBank and mapping databases such as the Genome Data Base in their research and to submit new results to be incorporated into the databases.

From a computing-oriented perspective, genome analysis is a data-driven science in which researchers must search massive amounts of data for the specific information they need to interpret their experimental results and plan new experiments. The unique product of a gene is a functional protein (such as an enzyme or a structural complex). Many more genes are known than the corresponding proteins; it takes person-years of effort to purify a protein, to understand its function, or to determine its three-dimensional structure. Such research conducted on the current scale would be nearly impossible without computing and information technology.

Such recent major efforts as the Human Genome Project, which seeks to map[3] and sequence all human genes, promise to generate data on a scale unprecedented in the history of molecular biology.

Using databases to store, access, compare, and analyze genetic data will thus become even more essential in the future than it is at present. In particular, the rapidly expanding body of sequence data makes it increasingly valuable to analyze the data effectively for functional patterns. There are already striking examples of how sequence analysis has provided insight and stimulated experimentation that would otherwise not have been an obvious line of investigation.

TRENDS IN TECHNOLOGY TO SUPPORT COLLABORATIVE GENOME RESEARCH

The anticipated avalanche of genome information makes it imperative that researchers have tools that will support using the data effectively to conduct research (see Box 4.2). Facilitating the capability to predict new relationships from existing data is an area ripe for the application of computer technology (Lander et al., 1991). Applications so far have concentrated on database management and on simple pattern analysis to enable homology searches. In contrast to space physics and oceanography, modeling and simulation have not been applied extensively in genome analysis.

Given that data from many laboratories are typically required to determine, for example, the part a gene may play in a particular function, communication and some degree of collaboration are necessary among genome researchers. Before computers were used, data were gathered from the literature and by direct contact with colleagues in other laboratories. Today, however, the shared knowledge necessary to achieve progress in genome research is often transferred via communal databases and other electronic modes of dissemination.

The traditional method for sharing new knowledge with other scientists has been to publish articles in the journal literature. In genome-related research, concise presentation of experimental results in fast-publication literature is emphasized. A typical article is 6 pages long and, if accepted, is published 3 to 6 months after submission. In addition to the results published in the traditional journal literature, the results of an experiment may include map or sequence data that are contributed to electronic databases.[4] The primary advantages of publishing in electronic format are that it is incrementally less expensive than publication in journals, allows searching and keeping data current to be done more easily, and enables the use of computer-based analysis programs.

In the foreseeable future, advances in support for collaborative genome analysis will likely come from enhancing scientists' interaction with databases. Integrated systems will not only store and retrieve data, but are also expected to facilitate the analysis of data and literature.

CURRENT DATABASE SYSTEMS

The computer and communications support for genome research focuses on networked databases for the submission, editing, retrieval, and analysis of data. Many databases have become essential tools for genome researchers, a status reflected in the creation of national facilities to maintain them. At present, the most important of the standard archival databases each contain a specific type of information. Historically, literature databases (e.g., MedLine) were generated first, followed by data archives for sequences (e.g., GenBank) and for maps (e.g., Genome Data Base).

These archival databases are becoming a new kind of scientific literature, and each has been faced with implementing aspects of the publishing process in electronic form. MedLine developed methods for electronic searching and retrieval of bibliographic information, GenBank developed methods for direct electronic data submission, and the Genome Data Base developed techniques for electronic checking of data by individuals at workstations distributed literally around the world.

Each of these projects costs several million dollars per year to operate. Although each is based at least in part on commercial database technology, all have required development of specific application software running to the hundreds of thousands of lines of code. The costs of operating these databases

reflect what is necessary to develop, maintain, and distribute the application software, to encode and check the data, and to provide appropriate user support and documentation. These databases afford economies of scale for genome researchers.

MedLine

The National Library of Medicine (NLM), a part of the National Institutes of Health, is the world's largest special-collections library. Its primary electronic collection is MedLine, which includes abstracts and bibliographic citations, and permits on-line searching of electronic versions of a wide range of journal articles of interest to practitioners and researchers in medicine and biology. Developed in the 1960s when computing and communications technology first began to enable remote information retrieval, MedLine is now accessed by essentially every molecular biology laboratory in the country as a matter of course, either through a centralized library subscription or an individual laboratory subscription.

MedLine currently includes over 6 million citations and is expanding rapidly; in 1990, for example, some 400,000 citations were added from 3,600 journals selected by an advisory committee. A group of professional indexers examines selected issues and generates citations for each article of biomedical interest (the abstract from the printed article, bibliographic information including author, title, journal, and volume and page numbers, plus classification information such as relevant terms from the standard NLM Medical Subject Heading (MeSH) thesaurus). These citations are then entered into the MedLine database.

For genome researchers and other scientists from all over the country who can routinely search the MedLine database, MedLine is a familiar example of an interactive database program with a front end running on their laboratory computers and the back end on a remote machine across the network. The front end is a simple text-based interface, the network is the telephone network accessed with a modem, and the back end provides simple Boolean word searching. Recently, MedLine has been made accessible via Internet. The MedLine database is also made available by other on-line search services such as Dialog and Bibliographic Retrieval Service. With the advent of CD-ROMs, many libraries and even some individual laboratories have purchased readers to run retrievals locally and save on-line charges. Typically, these collections are updated monthly rather than daily.

GenBank

Use of the centralized database GenBank has become routine for genome researchers. Whenever a gene is isolated and sequenced, it is compared with the sequences in the database in the hope of identifying its function by finding another gene of known function and similar sequence. As MedLine's capability for on-line literature searches has become essential to researchers, so also are GenBank's on-line data now essential to scientists who map and sequence genes.

GenBank began in the 1970s as a collection of nucleic acid sequences gathered by physicist Walter Goad, who was interested in sources for testing pattern-analysis algorithms. By the late 1970s, the collection itself was proving useful to biologists for making functional comparisons, and several striking discoveries were made directly from the database (Burks, 1985). In 1982, the Los Alamos Sequence Library became GenBank, the national repository for nucleic acid sequence data. Official responsibility has now been transferred to the National Library of Medicine, where the database will be maintained by the National Center for Biotechnology Information (NCBI) in collaboration with Los Alamos.

Searching of the GenBank database rarely involves manually examining many sequences with a coarse search criterion; instead, an analysis program using some form of similarity matching is utilized. Because such analysis programs tend to be slow and the database is still relatively small, GenBank is

commonly reproduced in its entirety at each local site and the copy is updated weekly or even daily from the central archive. At many large universities, the central biotechnology computing facility maintains a copy of GenBank that researchers can access from a computer in their laboratory over the campus local area network.

Initially, GenBank entries were generated in much the same way as MedLine entries. A group of indexers read the published literature, identified articles containing sequence data, and entered items into the central database. An item was like a citation; it consisted of a sequence plus a variety of annotations, including author, organism, methods, and possible functions. The central annotators typed in the sequence and entered all annotations, an approach that worked well for the first few years. However, the extraordinary growth in the number of sequences generated soon made this strategy less attractive, owing partly to the difficulty of accurately entering a large volume of hard-to-read data and partly to the changing economics of publishing whereby journals began to restrict the amount of sequence data that could be published.

Essentially all submissions to GenBank are now made electronically (Box 4.3). A program available at no cost to scientists provides an interactive form for entering the sequence data and the appropriate annotations and then electronically mailing this information to the central archive. Individual researchers as well as large-scale sequencing projects submit data to GenBank, and essentially every sequence submitted is incorporated into the database. Supporting automated submission has greatly increased the currency and accuracy of the database. In addition, although they exercise no editorial control, GenBank's curators run a series of tests (increasingly automated) on the data to check for accuracy, and they contact the authors when anomalies are found. The combination of electronic submission, which tends to eliminate transcription errors, and electronic checking, which tends to catch further anomalies, has significantly increased the level of review that data receive before being made public.

BOX 4.3 ELECTRONIC DATA PUBLISHING AND GENBANK

From a review article by the administrators of GenBank, the national repository for DNA sequence data: "Several years ago, in an effort to keep up with an exponentially increasing flow of nucleotide sequence data, we began looking at ways of combining two evolving technologies (namely, database management software and computer networks) into a more effective system for the communication of scientific data. We have designed and implemented a specialized form of electronic publishing, which we term electronic data publishing, where data (in our case DNA sequences and related annotation) are gathered, processed, and distributed electronically. Rather than compete with traditional scientific publications, electronic data publishing is designed both to complement and to support printed publications. . . .

". . . At the practical level, electronic data publishing at GenBank has been quite successful. . . . At a more conceptual level, however, some larger issues bear discussion. In particular, we must consider questions such as the peer review (and thus the quality) of submitted data and the academic credit that will be associated with database entries. . . . Our experience indicates that electronic data publishing has actually resulted in a higher level of quality for data in GenBank and therefore also in the journals. . . .

". . . [H]aving journal editors encourage or require submission of data before publication of a paper has been met with far greater acceptance on the part of authors than might have been anticipated. We are currently receiving approximately 80% of our data as a direct electronic submission before publication of any related paper. The research community will always decide de facto the relative importance of the generation of data; what we (and others who have acted as proponents of direct submission) have accomplished is the establishment of a system that researchers are coming to view as the natural mechanism for the communication of large amounts of scientific data."

SOURCE: Cinkosky et al. (1991), pp. 1273, 1276.

The establishment of the paradigm of electronic data publishing for GenBank illustrates some of the broader potential for collaboration technology. The most important aspect of the shift to electronic data publishing is recruiting of the community of researchers to perform the job of sequence entry and annotation and database building. This paradigm supports growth with resources that scale to the size of the community. The scientists who have determined the sequences are also the most appropriate people to annotate their own work, given a controlled semantics and syntax. The responsibility of the database managers is to specify the structure of the data entries and provide the community with the tools to generate entries according to these specifications. These tools also allow automatic checking of data, semantics, and syntax as well as writing a relational transaction for electronically entering data into the database.

Most molecular biology journals will not publish articles discussing sequences without a GenBank accession number, thus guaranteeing that the database is the medium of record. The GenBank staff worked closely with journal editors to develop policies requiring that data be submitted to a database before a related article could appear in print (Cinkosky et al., 1991).

The interval between discussion of a sequence in the journal literature and appearance of the sequence in the GenBank database has shrunk dramatically, from nearly 2 years to a few days. This reduction has paralleled an increase in the fraction of electronically submitted entries from 10 percent to about 90 percent. In 1993, the database will be 20 times the size it was 5 years ago. GenBank currently processes more sequence data in a month than it held in its entire database in 1987. This 10-fold increase in production was made possible by leveraging the efforts of the entire genome research community.

Electronic data publishing has also simplified collaboration with subscribers to other international databanks (e.g., the European Molecular Biology Laboratory Data Library and the Database of Japan, each of which exchanges new entries daily via the Internet. The currency of its database has made the GenBank On-Line Service a very valuable research tool, and daily GenBank updates are the critical reason that many molecular biologists have become part of the Internet community. Hence GenBank, a collaborative resource, has stimulated yet further collaboration.

The GenBank experience has shown that, with appropriate technological support and sufficient sociological incentives (such as the requirement by journals that sequences referenced in papers be deposited in a public database), the biological community itself can become actively involved in submitting information to and retrieving it from an archival database. The Human Genome Project promises to greatly increase the scale not only of data generated but also of scientists involved in direct maintenance of electronic archives. Computer support for a larger portion of the publishing process, in particular, editing and checking functions for quality control, can enhance the currency and accuracy of data made available to the molecular biology community.

Genome Data Base

In its initial phases, the Human Genome Project is concentrating on the sequencing and mapping of model-organism and human genes. The Genome Data Base (GDB) has been designated as the central database for human gene mapping data. Closely associated with the Welch Medical Library and maintained at the Johns Hopkins University Medical School, the GDB is operated in conjunction with the On-line Mendelian Inheritance in Man project, which provides ready access to detailed information about human genetic phenotypes and diseases.

Although the relational database underlying GDB contains more than 250 tables, GDB may be envisioned as containing data on four major classes of biomedical objects (map objects, proposed maps, mapping reagents, and clinical phenotypes) and two classes of supporting objects (references and people). In addition, many dependent object classes are also represented, such as alleles, polymorphisms, and populational allele frequencies. At the moment all of the data are stored in a commercial relational database, as with GenBank, and retrieved via simple text forms. It is expected that more sophisticated

graphical interfaces will become available, both from GDB and from other third-party developers, in the near future.

Data may be entered into the system through a variety of mechanisms, including direct submission by authors, entry by GDB staff, or input by editors or their assistants. The editors then guide the data through an approval process. Editorial control is explicitly supported by the software developed and maintained by the GDB project. A board of editors ensures that the database represents a consensus of checked mapping data. One or more editors act as referees for incoming data on each of the 24 human chromosomes. Other editorial groups address issues involving nomenclature, clinical phenotypes, and comparative mapping. Each editor has electronic access to GDB, which provides special interfaces and underlying functionality to facilitate editorial work. A complex system of approval flags and message files allows editorial groups to communicate with each other regarding aspects of the approval process for a particular piece of data. All editorial changes are logged, and users may call up and view at any time the editorial history of any entry.

Data transmission and interaction within this publishing process are supported electronically, although the actual decisions are made by the (human) editors. Such an environment enables the system to support the many laboratories that are major players in the genome mapping and sequencing project.

FUTURE INFORMATION SYSTEMS—TOWARD A WORKING COLLABORATORY

Steady progress has been made in increasing the functionality of computer support for collaborative research. MedLine, a product of the 1960s, provides remote interaction with a database generated at a central site. GenBank, a product of the 1980s, provides for electronic submission of data generated at laboratories across the international community and then sent to a central site for editing and redistribution. The Genome Data Base, a product of the 1990s, has developed electronic editing, with editorial activities also distributed and supported electronically.

To continue to provide increasing functionality to support collaboration in genome research, future information systems must address a number of needs. First, since an abundance of data is now available on-line, ways are needed to conveniently pass selections in and out of a variety of analysis programs. Second, since many different sources of data are now available, ways of establishing links between related items in different databases are required. Finally, a complete cycle of electronic publishing—from entry to editing to distribution of data—is needed not only for formal archival material but also for the informal community material that is also vital to science. In a sense, such a cycle brings the control of data generation, curatorship, and distribution directly to the scientific community. The next generation of support for collaborative genome research will include comprehensive computer environments for analyzing data. Such systems will additionally begin to support computational biology, in which substantial parts of some experiments can be done via computer programs that access databases and apply analysis packages.

The first step in supporting collaboration is to invoke analysis remotely across a network, as is done in literature searches that retrieve from MedLine abstracts that contain specified keywords.

The next step is to support analysis across multiple sources. Each data source represents a different experimental method, point of view, or level of knowledge. Interrelating these sources is an essential component of genome research, largely because the available knowledge is now so incomplete. To obtain as much information as possible about the biological function of the genes they are studying, researchers consult and interrelate gene lists, genetic maps, DNA sequence data, formal and informal literature, and other sources of information. Often they need to cross-compare among organisms, since different organisms have become models for studying different functions. Interconnecting different sources is accomplished by making associations between similar items from different databases. For example, standardized nomenclature, such as gene names, can be used as keys to associate aspects of the

genetic maps of different organisms. In a more generalized sense, searches that discover semantic similarities can be used to associate related items.

The final step in the next generation of support for collaboration in genome research is to support a complete system for the genome research community that combines the association of information from multiple sources provided by similarity analysis software with the range of knowledge provided by a complete electronic publishing environment. Such a community system, or collaboratory, will capture all of the specialized knowledge needed by a geographically distributed community of researchers, including informal, more detailed sources of information, and the quality-assurance mechanisms needed to check these items. Scientists will then interact transparently with this knowledge across the national networks, both retrieving and analyzing existing information, as well as submitting and publishing new knowledge. The goal is to extend the publishing process into a distributed electronic medium to enable the whole community of genome researchers to contribute to and edit its knowledge. This functionality is available in research prototypes discussed below.

Basic Components and Research Prototypes

Remote Analysis Servers—Blast

To implement a complete analysis environment, it is first necessary to support the transparent application of analysis programs. A remote analysis server enables scientists to use large, remote computers to analyze their data; on-line use of MedLine provides a simple version of such a server. A version with modern technology is commonly used at the National Science Foundation's supercomputer centers to enable scientific visualization: an interactive graphical front end runs on a personal computer to enable the user to issue commands and display results; this is connected via the Internet to a powerful back-end supercomputer running a sophisticated, computationally intensive numerical simulation.

A research version of a remote analysis server for genome research is the "Blast" server maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. "Blast" is one of the best current sequence analysis programs (Altschul, 1990); given a sequence, it produces a rank-ordered list of the sequences in the database that are most similar under a set of biological heuristics. NCBI has set up a continuously running server for Blast on its fast parallel machine, using the daily updated GenBank database. An Internet electronic mail client enables biologists to submit their sequences, have them mailed electronically to NCBI, and have the results mailed back, usually in a matter of minutes.

A good illustration of how scientific collaboration can be facilitated by computer technology is the expressed sequence tag (EST) project (Box 4.4), which has the Blast server as an essential component (Adams, 1991; Fields, 1992).

Interconnecting Archival Databases—Entrez

A complete analysis environment must enable analysis of the similarity of relationships among items not only within a single database, but also across multiple databases. To ensure that multiple sources can be easily interlinked, it is necessary to define a standard format for passing items between analysis programs.

To facilitate interconnection between the literature database MedLine and the sequence database GenBank, indexers (human librarians) at the National Library of Medicine are placing standard links between related items in the two databases. A field for every MedLine entry specifies referenced sequences, and a field for every GenBank entry specifies referenced literature. The professional indexers

BOX 4.4 A NEW COLLABORATION EVERY DAY: IDENTIFYING HUMAN GENES USING AUTOMATED SEQUENCERS AND THE INTERNET

Cloning, sequencing, and characterizing the biological function of a gene are long, slow processes requiring many different types of biological expertise. In 1990, J. Craig Venter at the National Institute of Neurological Disorders and Stroke developed a shortcut method for identifying genes: using automated DNA sequencers to sequence a few hundred bases from one end of a complementary DNA (cDNA) clone of the gene's messenger RNA (mRNA) transcript. The resulting "expressed sequence tag" (EST) is a unique identifier for the gene. ESTs generated from randomly chosen cDNA clones effectively provide a clone and a partial sequence of a gene in a single experiment. In many cases, the gene family to which a newly cloned gene belongs can be identified by comparing the EST sequence to the known sequences contained in databases such as GenBank. The National Center for Biotechnology Information's "Blast" Internet server and Oak Ridge National Laboratory's GRAIL server are key resources in this analysis process. In the last 2 years, the EST project has published sequences of thousands of human genes, developed a database of sequences and analysis results that is available over the Internet, and provided the clones from which these ESTs were derived to the research community via the American Type Culture Collection. Venter's laboratory, now at the Institute for Genomic Research, is sequencing thousands of new ESTs per month.

Each EST raises the possibility of a new scientific collaboration to characterize the function and assess the biological and medical significance of a new human gene, and hundreds of requests for sequences, data, clones, or additional information have been answered by the laboratory. In some cases, members of Venter's laboratory participate in a collaborative follow-up analysis, doing additional sequencing or other experiments. Several human genes related to the "discs-large" tumor suppressor gene in drosophilia, for example, are being pursued in collaboration with colleagues at the University of California, Irvine and San Francisco. In many cases, data and materials are sent for others to pursue independently. Either way, the rapid communication facilities provided by the Internet allow quantities of data that could not feasibly be published in print media to be exchanged quickly and efficiently between colleagues around the United States and in many other countries.

SOURCE: Adapted from Fields (1992).

specify these links when they specify classification and annotation information. Entries in both databases have unique identifiers so that the links can be uniquely and permanently specified. Once the links have been established, software can be written to enable retrieval of related items by following the relationship links. For standard archival sources, such explicit manual specification of interconnections promises great utility.

There are many archival sources in molecular biology, covering the many different experimental aspects of the field. Currently a few such sources, maintained by a single library, have an explicit set of maintained links. For other sources, the links must be maintained automatically or semi-automatically by attempting to match "common" items in different databases. For example, both DNA sequence and protein sequence databases often mention the corresponding gene name for a given organism. These common names could be used automatically as points of interconnection if there were a standard nomenclature for naming genes. These implicit (discovered) links will likely be far more numerous than the explicit (manual) ones.

Standardized formats are also necessary before name matching can be done. The fields representing the annotations of a sequence, for example, must be of the same set before their corresponding values can be matched. A program must be able to tell, for example, that the "gene name" field in a DNA sequence database is the same as the "function name" field in a protein sequence database. To describe data types in molecular biology, the National Center for Biotechnology Information (NCBI) is using the International Organization for Standardization standard language ASN.1 (Abstract Syntax Notation) to describe data fields and their values.

The NCBI ASN.1 toolkit provides a variety of parsing programs and a set of semantic definitions for the values for the most common types of genome data, such as literature and sequences (Gish et al., 1992). While ASN.1 provides a standard interface for the programs of the NCBI toolkit, it is not broadly implemented in other information processing environments in molecular biology. However, authors of analysis and display software are gradually adopting ASN.1 as their external publication format for data exchange, partly as a result of NCBI's stature within the community. Adoption of a standard interface for passing data between analysis servers would greatly facilitate development of a complete analysis environment. For example, any server that manipulates DNA sequence data should be able to pass and return the same representation for a sequence. Ideally, analysis environments should need to provide an invocation procedure only for each data type, rather than for each data server—a feature that is more feasible for genome data given that the number of data types is fairly small.

A good illustration of a software package utilizing this technology is Entrez (Gish et al., 1992), which currently supports interactive cross-comparison of data in MedLine and GenBank. The base data are the data in MedLine and GenBank and the relationship links explicitly established between entries in these two databases. The base software is the Blast server. Users can specify desired keywords and retrieve related documents from MedLine and can then follow related links from the MedLine documents to a set of GenBank sequences, run the sequence searcher to find a set of similar sequences, follow the related links from this new set back to a set of related literature documents, run the text searcher to find a set of related documents, and so on, thus engaging in an iterative process of pattern matching to make connections that could not be discovered using a single database analysis.

Systems like Entrez will soon include a wider variety of databases in molecular biology. For example, NCBI is already expanding Entrez to include protein databases. Increasing the number of databases in integrated environments increases the possibility of discovering important similarity matches or patterns that might not have been apparent otherwise.

Community Information Systems

A complete community information system to support research would encompass an electronic library containing all sources of data, software enabling interactive display and analysis of all data types, and an underlying information infrastructure that transparently supports extensive facilities for browsing, filtering, and sharing knowledge across the international Internet. Building working models of such a complete collaboration system is possible today and is actually being done in an ongoing research project, the Worm Community System.

Basic Elements and Capabilities

The first step in building a community information system is to decide the extent of the knowledge to be captured, which depends in turn on the needs of the particular community whose work is to be facilitated. Not all of the elements in the scientific process described in Chapter 1 are equally important in genome research. For example, investigators rarely wish to examine actual raw data and are satisfied if the maintainers of instruments, e.g., those in the large sequencing projects, keep the raw data streams and release only a final version to the archival databases. Because experiments in genome research typically involve only a few people or only loose cooperation between different laboratories, the need for real-time cooperative work tools is relatively small, and support for collaboration should focus on retrieval and analysis of the archival data and literature, and on extending the same facilities to more informal material that is critical to progress in science.

In genome research, this informal material typically includes results that are not yet publishable in journal articles but may be discussed in newsletters (e.g., giving one- to two-page descriptions of

current experiments) and in conference proceedings (e.g., giving abstracts of new results). It also typically includes data too specialized to be publishable in archival databases—such as strain lists (local sets of mutant organisms) and restriction maps (detailed locations of genes)—yet still extremely useful to other researchers working on similar problems. These informal results are usually stored in individual laboratories inside filing cabinets or taped to the backs of doors and are usually retrieved from the outside only by telephoning the laboratory. Electronic sharing of this knowledge would make it available to a much wider set of potential collaborators.

To collect and check this informal knowledge, the appropriate biological community must be directly involved. The experience with GenBank demonstrates that biologists conducting genome research will electronically submit material that is deemed useful, as does the extensive experience of scientists with electronic bulletin boards such as Netnews. The experience with the Genome Data Base shows that electronic support for distributed editors (human editors to check the quality of materials) is workable with community-chosen curators of the individual data sources. A complete electronic publishing environment will support entry of all the specific types of knowledge and publishing in all the different styles. With such a system, for example, an investigator will be able to run a program that enables interactive specification of a restriction map or preparation of a newsletter article and then automatically submits it to a central archive, which then automatically distributes it to the community. Local maps might be unchecked and newsletter articles might be moderated (checked for topic), as opposed to centrally archived data or literature that must be checked for quality and consistency.

The process of associating and relating items from myriad formal and informal sources involves major technological and sociological complications. Interconnection of informal sources of information depends on implicit associations being made automatically, e.g., by parsing the text of newsletter articles to locate gene names with which to make associations. In a community information system, interconnection is facilitated by community members themselves, who may add most associations. That is, the users of the system can specify their own associations between items, including associations discovered while using the system itself.

A distributed system is required to support sharing of information in a scientific community, since the users and the generators of knowledge are geographically distributed. National networks are expected to be a necessary component of a community information system or analysis environment, even if the total amount of knowledge is small. A community information system should present users with what appears to be a single logical "database" for retrieval and storage of data.

A Model Collaboratory—Worm Community System

A model of a complete community information system is under active development in the Community Systems Laboratory at the University of Arizona (Box 4.5). This system represents a substantial operational model of a collaboratory. The community is the collection of molecular biologists who study the nematode worm *Caenorhabditis elegans*; the system itself is called the Worm Community System. The size of the community and the amount of data are manageable—large enough to be interesting from the perspective of data organization and management yet small enough to be doable. The community's approximately 500 members are spread across the United States, Europe, and Japan. Significant sources of data already exist in electronic format, with a range of types and location. Community members have a long tradition of openness and sharing of data.

The underlying technology of a community information system such as WCS is based on a representation called an information space (Schatz, 1987, 1989) that supports transparent manipulation of objects from multiple, heterogeneous, distributed sources and is thus a form of a federated object-oriented database. The goal is to enable users to manipulate data items from many sources of different types as though the items are uniform units of information. The information space consists of the set of associations among these information units. Thus, a single set of user commands suffices to browse,

filter, and share all the different types of data. This is accomplished internally by packaging data from all external sources as uniform objects with a generic set of operations for publishing, searching, displaying, and associating the data. The system itself also handles retrieval of objects from remote sources across the network and provides the necessary caching policies to make this retrieval speed-transparent. For the types of data common in genome research, the existing Internet has sufficient bandwidth to support interactive retrieval across the country.

BOX 4.5 A MODEL COLLABORATORY

Genome research is the subject domain for a current example of a functioning national collaboratory. The Worm Community System comprises a digital library containing the data of the community of molecular biologists who study the nematode worm *Caenorhabditis elegans*, which has become a primary model organism in the Human Genome Project, and a software environment that supports interactive manipulation of this library across the Internet. The current library contains a substantial fraction of the extensive knowledge about the worm, including gene descriptions, genetic maps, physical maps, DNA sequences, formal journal literature, informal newsletter literature, and a wide variety of other informal materials. The current environment enables users to browse the library by search and navigation, to examine and analyze selected materials, and then to share composed 'hyperdocuments' within the community. The current prototype is running in some 25 worm laboratories nationwide, and there are already instances of users electronically submitting items to the 'central' information space and having these automatically redistributed to other sites. The next release of the system will support electronic publication with editorial levels, as well as invocation of external analysis programs. Subsequent releases will move toward a complete analysis environment, with a large collection of databases and literature accessible transparently across the national network for examination and for analysis.

The first release of the Worm Community System is now running in worm laboratories across the country and supports a sample range of the necessary knowledge and functionality (Figure 4.1). The second release will be available in 1993 and will support a sample range of publishing mechanisms. As the publishing system and the electronic community evolve, subsequent releases will support deeper knowledge semantics and begin to move toward a generic information infrastructure.

OPPORTUNITIES TO ENHANCE RESEARCH

Genome research in molecular biology has undergone a significant revolution owing to the existence of archival databases such as MedLine, GenBank, and the Genome Data Base. All practicing genome researchers consider it essential to cross-compare their experimental results with existing results by running similarity searches on these databases. Next-generation central archives, now in use in research prototypes, promise even greater utility and opportunities for collaboration. Remote analysis servers (e.g., Blast) will enable rapid, daily comparisons of sequences. Systems for interconnecting multiple archives (e.g., Entrez) will enable rapid comparison across different sources. The fact that such prototypes are being implemented with a standard data exchange format by the National Center for Biotechnology Information at the National Library of Medicine promises that integrated analysis environments for standard archives will become a reality in the foreseeable future.

At the same time, models for a complete collaboratory are also being developed. The Worm Community System illustrates what a complete collaboratory in genome research could become in the foreseeable future. It provides analysis of both formal and informal knowledge and electronic sharing of user-provided knowledge. As a distributed system utilizing the existing network communications infrastructure, it points the way toward a national information infrastructure that will enable scientists to

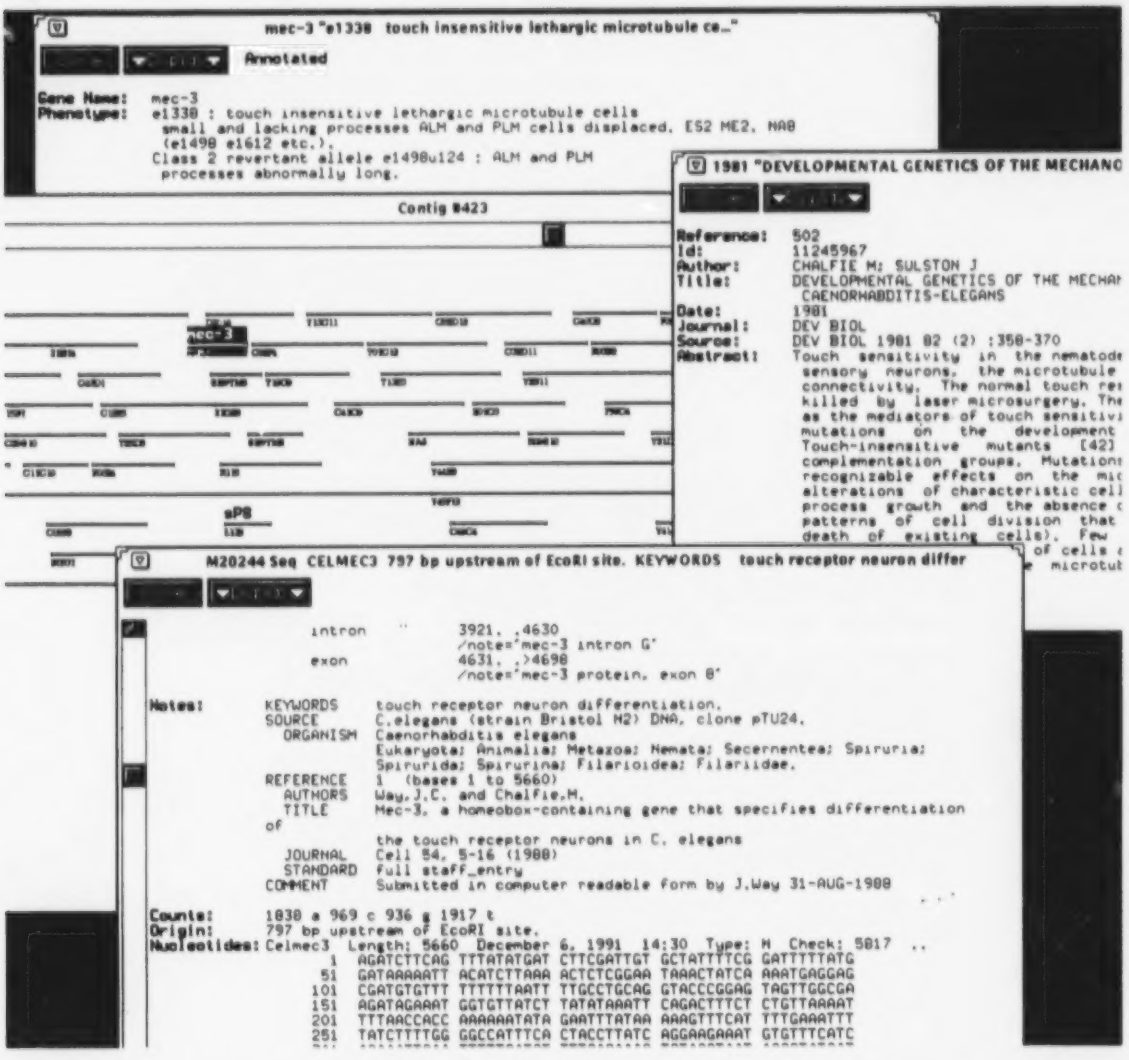

mec-3 "e1338 touch insensitive lethargic microtubule ce..."
Annotated
Gene Name: mec-3
Phenotype: e1338 : touch insensitive lethargic microtubule cells
small and lacking processes ALM and PLM cells displaced. ES2 ME2, MA8
(e1498 e1612 etc.).
Class 2 revertant allele e1498u124 : ALM and PLM
processes abnormally long.
Contig 8423
mec-3
sPS
1981 "DEVELOPMENTAL GENETICS OF THE MECHANO
Reference: 502
Id: 11245967
Author: CHALFIE M; SULSTON J
Title: DEVELOPMENTAL GENETICS OF THE MECHAN
CAENORHABDITIS-ELEGANS
Date: 1981
Journal: DEV BIOL
Source: DEV BIOL 1981 82 (2) :350-370
Abstract: Touch sensitivity in the nematode
sensory neurons, the microtubule
connectivity. The normal touch re
killed by laser microsurgery. The
as the mediators of touch sensitivi
mutations on the development
Touch-insensitive mutants [42]
complementation groups. Mutation
recognizable effects on the mic
alterations of characteristic cell
process growth and the absence
patterns of cell division that
death of existing cells). Few
of cells
microtub
M20244 Seq CELMEC3 797 bp upstream of EcoRI site. KEYWORDS touch receptor neuron differ
intron 3921. .4630
/note="mec-3 intron G"
exon 4631. .>4698
/note="mec-3 protein, exon 8"
Notes: KEYWORDS touch receptor neuron differentiation.
SOURCE C.elegans (strain Bristol M2) DNA, clone pTU24.
ORGANISM Caenorhabditis elegans
Eukaryota; Animalia; Metazoa; Nemata; Secernentea; Spiruria;
Spirurida; Spirurina; Filarioidea; Filariidae.
REFERENCE 1 (bases 1 to 5660)
AUTHORS Way,J.C. and Chalfie,M.
TITLE Mec-3, a homeobox-containing gene that specifies differentiation
of
the touch receptor neurons in C. elegans
JOURNAL Cell 54, 5-16 (1988)
STANDARD full staff_entry
COMMENT Submitted in computer readable form by J.Way 31-AUG-1988
Counts: 1838 a 969 c 936 g 1917 t
Origin: 797 bp upstream of EcoRI site.
Nucleotides: Celmec3 Length: 5660 December 6, 1991 14:30 Type: N Check: 5817 ..
1 AGATCTTCAG TTTATATGAT CTTCGATTGT GCTATTTTCG GATTTTTATG
51 GATAAAAATT ACATCTTAAA ACTCTCGGAA TAAACTATCA AAATGAGGAG
101 CGATGTGTTT TTTTTTAATT TTGCCTGCAG GTACCCGGAG TAGTTGGCGA
151 AGATAGAAAT GGTGTTATCT TATATAAATT CAGACTTTCT CTGTTAAAAT
201 TTTAACCACC AAAAAATATA GAATTTATAA AAAGTTTCAT TTTGAAATTT
251 TATCTTTTGG GGCCATTTCA CTACCTTATC AGGAAGAAAT GTGTTTCATC

FIGURE 4.1 Sample session with release 1 of the Worm Community System, illustrating what might occur when a molecular biologist interacts with the community library. Shown are the coverage of both data and literature, and some of the relationship links. The user began with a broad query of the term "sensory," which returned all items from all sources mentioning that term, including the formal literature, informal literature, gene descriptions, sequence annotations, and so on. By browsing through short summaries of these items, the user found a literature item describing a number of mechanosensory genes (shown on the right). The relationship links to this literature article were then followed to retrieve a set of gene descriptions. The gene "mec-3" was of particular interest, as shown at the top left. From this gene description, the physical map was selected and an interactive display of the DNA clones appeared centered around where the gene was located (shown in middle left). This graphical display can be selected and manipulated; in this case a further zoom or link following was done to retrieve further information, which included the DNA sequence shown in the bottom left. Note that each of the items shown (literature, gene, map, sequence) comes originally from its own database, but the community information system enables navigation across all these sources with single, uniform commands. Not shown is a further interaction made possible by using an analysis program on the sequence to display its coding regions. SOURCE: Courtesy of Bruce Schatz, University of Arizona.

manipulate sources of information transparently across the country. However, it is still a preliminary model and will have to be expanded considerably to demonstrate a functional community information infrastructure. For example, the sets of knowledge and analysis must be expanded, a true distributed system across platforms and networks must be developed, and a case-hardened implementation must be evolved before the technology is ready to support standard archives. But with sufficient resources, complete collaboratories for sharing, comparing, and analyzing data will be built for genome research since the need is there and the technology is available. The pattern discovery enabled by "dry-lab" analysis environments promises another significant revolution for the support of research in molecular biology.

NOTES

1. Replication is the process in which existing DNA is used as a template for the synthesis of new DNA strands. Mutagenesis is the process by which DNA is mutated or modified. Transcription is the synthesis of RNA, a long-chain nucleic acid consisting of repeating nucleotide units, from a sequence of DNA. Translation is the process in which the genetic code directs the synthesis of proteins from amino acids.

2. Two examples of gene sequencing technology are Polymerase Chain Reaction (PCR) and gel electrophoresis. PCR is a method for increasing the number of copies of a specific DNA fragment to make the fragment easier to detect and identify. Gel electrophoresis is a method of separating large molecules in an electric field, allowing DNA fragments differing by single bases to be readily separated. Combined with methods such as Sanger's dideoxynucleotide chain termination procedure, gel electrophoresis can produce ladders of DNA molecules from which DNA sequences can be determined.

3. "A primary goal of the Human Genome Project is to make a series of descriptive diagrams—maps—of each human chromosome at increasingly finer resolutions. Mapping involves (1) dividing the chromosomes into smaller fragments that can be propagated and characterized and (2) ordering (mapping) them to correspond to their respective locations on the chromosomes. After mapping is completed, the next step is to determine the sequence of base pairs of the ordered DNA fragments. A genome map describes the order of genes or other markers and the spacing between them on each chromosome. If the full sequence of genes were known, research emphasis could shift to determining gene function." (Cantor and Spengler, 1992, p. 198)

4. The common stages for publishing data in the electronic domain are similar to the steps in ordinary publishing. The raw data are recorded in laboratory notebooks that are kept private, or kept on archival disk if generated directly by a sequencing machine. A processed form of these data, such as a map location or a sequence, is submitted for inclusion into a database. This is edited for publication by a central editor; typically a single curator chooses what data in what form will be included. Finally, the edited database is distributed for use by other biologists.

5

Building and Using Collaboratories

The promise of collaboratories is that—if they are thoughtfully developed to meet the needs of working scientists for handling information in its many forms—they have the potential to free researchers to concentrate on the purpose and results, rather than the mechanics, of communicating. If, in time, interactive "center[s] without walls" become a reality, then such collaboratories may further contribute to a positive and reinforcing sense of community among scientists that increases as the scale, quality, and scope of the shared information grow.

Building collaboratories that can cost-effectively facilitate scientific research requires that several technical, social, organizational, and practical issues be recognized and dealt with. Some of these, in many cases representing a distillation of discipline-specific needs and problems pointed out in Chapters 2 through 4, are outlined below.

IDENTIFYING BASIC CAPABILITIES A COLLABORATORY SHOULD SUPPORT

Of the many capabilities a collaboratory might be envisioned as providing, the following four classes address common information-related problems that have led, for the most part, to ad hoc and idiosyncratic solutions. Collaboratories would integrate some or all of these capabilities (depending on the needs of the relevant scientists). They would also foster development of better tools.

1. *Data sharing*. The capability for scientists in different locations working on the same project to quickly and easily obtain access to data, both within and across databases.

2. *Software sharing*. The capability for scientists in different locations to conveniently share software that supports data analysis, visualization, and modeling.

3. *Controlling remote instruments*. The capability for scientists to control instruments located in difficult-to-access regions on Earth, or in space, for example.

4. *Communicating with remote colleagues*. The capability for scientists in different locations to interact effectively with one another despite separation in space and/or time.

For these general classes of problems, common technologies that provide the desired capabilities and thus could be broadly useful across scientific disciplines might be made available through a public infrastructure (such as that provided today by the Internet, which enables sending messages and transferring files); in the form of general-purpose tools; or as a substrate for more specialized tools (as in the case of remote instrument control). Tools specific to particular disciplines, types of research problems, or projects will always be necessary, and they will have to be developed within the context of

individual projects or disciplines. However, more widely applicable tools and infrastructure might be the initial components of collaboratories designed to support doing science in collaboration at a distance.

PROVIDING BASIC CAPABILITIES—TECHNICAL CONSIDERATIONS

Today's computing and communications infrastructure supports rudimentary collaboration at a distance but typically is inadequately developed, deployed, and supported to sustain the quality and scope of tools and applications envisioned for collaboratories that will facilitate scientific research.

Interconnecting Data Sources

Scientists participating in this project's three workshops discussed a number of components considered essential to an enhanced capability for sharing data:

- Electronic libraries that would combine databases, literature, and software relevant to their research. Scientists considered electronic libraries a top priority because of the potential for rapid access to literature and the library's capacity to help locate information and data.

- Easily accessible archives of data, particularly in the physical sciences, for some of which large, established archives already exist, such as those at the National Space Science Data Center and the National Center for Atmospheric Research. Archives have become increasingly important as experiments and data gathering have become more complex and expensive, and as data sets have become more massive.

- A comprehensive system that would support retrieval of data from any or all sources, regardless of the data's origin or physical location. An example is the globe data catalog contemplated in Chapter 2, which would visually relate collected and archived data to the area of investigation, the type of data collected, and the time period that the data derive from. Today, data catalogs of archived holdings do exist, but the format for their presentation, the means of searching for the desired information, and the accessibility of the catalogs to researchers all vary widely from archive to archive, especially across disciplines.

Currently, research prototypes exist that permit users to issue single queries that search across multiple databases or archives. One example is the Worm Community System (Schatz, 1991-1992), described in Chapter 4, which represents a specific solution to a small research community's requirements for sharing data. Another useful tool is resource discovery software, which searches descriptions of databases or files to locate suitable sources to search and which can transfer files once they are discovered. Widely used examples of public-domain software include Archie,[1] Gopher,[2] and World-Wide Web,[3] which search by name of the file, and Wide Area Information Server,[4] which permits free-text searches by concept. These resource discovery tools were designed to be used on the Internet—an environment where it is expected that files will be shared. In a general-purpose file system—an environment in which the sharing of files may be an afterthought to their creation—resource discovery is much more difficult. Nevertheless, prototype resource discovery tools are now being developed for general-purpose file systems. One such prototype is Essence (Hardy and Schwartz, 1993). Despite current research efforts, more work needs to be done in this area.

If all or nearly all the information sources for a given subject domain can be accessed uniformly, the resulting ensemble of all information sources becomes a powerful research tool. The entire corpus of sources can be considered a single federated database consisting of multiple physical databases and

often containing different types of data, but giving the appearance of a single, logical whole. Supporting the appearance of uniform retrieval across data sources requires standard protocol interfaces and transformation programs to change the representation of each type of data into a standard format. Transformations such as those for text, graphics, and image conversions could be generic across all disciplines. Others would be subject-matter specific, such as those for maps and sequences for genome research or for temperature and currents for physical oceanography. Although the external data formats can vary considerably, federation across data types is possible with standard formats for representing the internal data. In standard database technology the standard query language (SQL) and Open-SQL interfaces provide a fundamental part of this linkage, but other structures for mapping data dictionaries and semantic values must also be developed for a federated database to be constructed.

For sources containing textual (and other) information, some progress has been made in standardizing information search and retrieval protocols. The recently developed American National Standard Z39.50, Information and Retrieval Service Definition and Protocol Specifications for Library Applications, provides the means for performing queries on textual information and is being adapted by the International Organization for Standardization as an international standard. However, it is only one standard with a modest number of applications and a multitude of data formats to search across. Consequently, while the Z39.50 standard is a good start, much more needs to be done to extend this protocol and to further develop other appropriate standards and protocols for system-independent data search and information interchange.

In addition to conducting broad searches across many databases, scientists may wish to record logical associations they detect between items within a database or across databases. Such associative links may build on previously identified relationships or represent the exploration of new ones among the database elements. For example, genes might be represented by linking various elements in gene map and sequence databases. An ocean voyage might be represented by a set of linked oceanographic database items. Unusual and nonintuitive links among database items recorded by one researcher may well stimulate new insights or approaches to a problem by other scientists.

Implementing logical links between related items in different sources requires a standard format for representing the links and a series of methods for determining semantic relationships.[5] These are areas for research and development. The commonest method for identifying semantic relationships relies on the use of standard terms or nomenclature such as the well-defined names that denote particular genes described in the literature, maps, and sequences. When standard nomenclature or terms have been used, it is possible to automatically generate links. For less obvious or novel relationships, a collaboratory system that supports data sharing should be able to support user-specified links.

Sharing and Applying Programs

Analysis of collected data lies at the heart of the scientific process. Data to be analyzed can be numeric or symbolic; some data, for example, may be in the form of literature. Increasingly, scientific analysis involves the use of software. Currently, genome researchers use analysis software to locate related items, such as gene sequences similar to a sequence being studied. In physical oceanography, simulation software is used to predict the results of future experiments, such as projecting ocean currents in a particular region. In space physics, modeling software is used to predict the behavior of observed phenomena, such as effects of the solar wind on the aurora. Community software such as IRAF and AIPS in the astronomy community and ORTEP and X-PLOR in the molecular biology community have proven very useful and have been widely disseminated and shared.

The committee found that sharing of software, application of external (i.e., not local) software to data, and application of local software to external data were three important capabilities sought by scientists contemplating useful collaboratory tools. Workshop participants observed that their research would be facilitated by technology that would allow them to call their specialized programs into action

easily and consistently, operate on data retrieved from their own and other data sources, and store results in network-accessible archives.

Sharing of software is becoming increasingly attractive to scientists as the sophistication of software—for example, visualization tools for displaying complex, multidimensional data—grows and as the investment required to develop complicated programs increases. One vehicle for sharing scientific software is the supercomputer centers: the National Center for Supercomputing Applications and the San Diego Supercomputer Center, for example, have become visualization centers—in part because of their major investments in high-performance computing hardware, software, and skills that can be leveraged by scientists in a variety of fields. These centers demonstrate how user/scientists can partner with developer/technologists to the benefit of both. Scientists and funders of research, such as the National Science Foundation, recognize that adapting a tool developed for one discipline (or project) for use in another may be less expensive than developing a new tool from scratch.

Despite its attractiveness, software sharing may be easier said than done. Users of borrowed programs may need to know the specifics of the numeric methods applied in an analysis tool. The proper interpretation of results may be greatly influenced by the choice of computational method and particular implementation of the software on a particular platform. These observations suggest that documenting program functionality, operation, and implementation, at least for programs made available for third-party use, is just as important as documenting the circumstances under which data may have been collected, so as to guide future analysis. In the absence of adequate funding for information tool development and dissemination, documentation of home-grown tools has tended to be limited to nonexistent, contributing to the tendency for scientists to replicate efforts. By explicitly underwriting the cost of tool development and dissemination, a collaboratory initiative could help to assure that appropriate documentation is developed and made available, as well as support the development of easier-to-use software.

Applying software that may not be collocated with the data of interest is difficult today, because most scientific software is prepared for specific purposes, and the data with which it is used must be carefully formatted to match. Each program has its own calling sequences[6] and user interface; there is little uniformity. For some applications such as statistical analysis, libraries of subroutines have been developed to perform commonly needed computations, but even when these routines are used, data must be formatted and passed in accordance with conventions or standards established for the library routines. In a collaboratory system, applying remote, network-based software would require a calling convention that supports remote program execution with standards for passing typed objects back and forth. Research and development are needed to create better conventions, which may in turn be adapted as standards.[7]

Some of the capabilities desired by scientists in wide-area, networked environments are in evidence on the smaller, simpler, and local scale of personal computer (PC) systems. Multifunction PC software integrates database, word-processing, spreadsheet, and electronic mail packages; in other cases, PC software is designed for ease of transfer of data between one kind of application and another. For example, spreadsheet results can be represented in bar or pie chart form and inserted into a multifont, compound text document; spreadsheets may refer to database entries; and text may be merged with the contents of one or more databases. In general, such integrated systems must be developed around common format conventions supporting data exchange or conversion, a process that has proved to be easier in the PC environment than in the more demanding scientific computing environment. The development of multifunction PC software has also benefited from the existence of a large commercial market, and it is likely that commercially developed technology such as video conferencing, "groupware," and computer-supported cooperative work-tools will benefit the development of collaboratory systems for software sharing as well.

Implicit, but not explicitly addressed in the workshops, was the notion that improved algorithms would also benefit some scientists. Collaborations between computer scientists and other scientists could advance the state of algorithms applied in scientific research—one of the objectives of the High

Performance Computing and Communications initiative. However, some software is so specific to a given discipline that the scientists involved must develop it or participate heavily in its development.

Controlling Remote Instruments

The value of controlling instruments through a computer network and collecting data regardless of the instruments' location depends on both the inaccessibility of the instruments and the difficulty of collecting data in the given environment. Remote control of instruments and remote data collection are required capabilities for space physicists, for example, who must collect data from distant reaches of space either through ground-based instruments positioned in remote locations, such as the Sondre Stromfjord Observatory in Greenland, or through space-based instruments that may need to be retargeted during the progress of a mission. Collection of data from remote instruments is also of major importance to physical oceanographers, whose data-gathering buoys and moorings are widely dispersed across the entire ocean surface and are visited only infrequently by research ships. In oceanography, the ability to better perform real-time reading of remote instruments would save time and money and would likely result in a much greater volume of higher-quality data being collected. Remote control of instruments and collection of data from remote instruments are considered relatively unimportant for genome mapping and sequencing at this time, although that situation could change.

Remote instrument control requires reliable networking with sufficient speed to support interactive responses. Controlling an instrument across a network requires a method for capturing the output of the instrument and transmitting it to the researcher who is directing the instrument, and a method for gathering user commands and transmitting them as inputs to the instrument, so that, for example, the instrument can make a requested change in its data-gathering procedure. Assuming that the instrument can accept commands and respond to them to achieve the desired results, remote control can be implemented with an appropriate network connection. Consequently, there will be an increasing need for standards for telemetry for remote control of instruments. The space science community is pursuing the development of standards for remote instrument control through the international Consultative Committee for Space Data Systems (CCSDS), which includes representatives from the space agencies of 26 countries, including the United States, countries in the European Community, and Japan. The CCSDS has also been responsible for developing common standards for space-ground links, multiplexing of high- and low-speed data streams, and data formatting and authentication.

It is apparent that scientists are making increasing use of remotely controlled instruments in ways not possible before the advent of computer networking. One such instrument is an Internet-based electron microscope that delivers high-quality imagery over the network and in return allows remote positioning of the viewing stage (Box 5.1).

Similarly, some telescopes have been outfitted with network connections, so that they can deliver digitized views, captured in high-resolution charge-coupled-device arrays, essentially anywhere accessible by the Internet. Other sharable instruments could include particle accelerators and colliders, radio telescopes, various satellite or space-platform instruments, autonomous underwater vehicles, pilotless aircraft, and autonomous land rovers.

Supporting User Interaction

Collaboration requires cooperation, and cooperation implies communication. An essential component of a collaboratory is the capability to support user interaction ranging from immediate, real-time, face-to-face discussion to deferred informal messaging, and even formal exchange of refereed papers. The support of interpersonal interaction among a group of collaborators may be the most chal-

BOX 5.1 DIAL-A-MICROSCOPE

"Want to access an electron microscope without leaving the comfort of your own office? Soon you may be able to log onto a computer network and start collecting images . . . on the 400,000-volt electron microscope at the University of California, San Diego.

"The concept is called the Microscopist's Workstation, which . . . enables a researcher with a computer workstation and access to Internet or NSFnet to control the microscope in real time. (A technician prepares the samples and puts them under the lens.) Project leader Mark Ellisman, a neuroscientist at the University of California, San Diego, unveiled his project at SIGGraph 92, an international conference on computer graphics held . . . in Chicago. One session, which included Ellisman's and 34 other projects, focused on how high-speed computer networks might bring the lab to the scientists: In addition to hopping on the La Jolla scope, attendees previewed hookups that would allow scientists to use their computers to walk through the internal organs of a 7-week-old embryo, go on a tumor safari in a human brain, and interact with a developing thunderstorm, with all the images generated on a remote supercomputer. Ellisman says, '. . . . This is just the first step in a long-term collaboration . . . to create a distributed laboratory that will make expensive national resources more widely available to the U.S. community.'"

SOURCE: "Dial-a-Microscope," *Science* (21 August 1992) 257:1048.

lenging aspect of collaboratory construction: it not only involves potentially all of the technical features of systems to access remote data, programs, and instruments, as well as multimedia work-group communication systems, but also requires an understanding of the complexities and vicissitudes of human behavior.

The most widely available technologies that support user interaction are electronic mail and facsimile transmission, which support asynchronous communication. Other applications include bulletin board systems, computer conferencing systems, file and document storage and retrieval systems, and the relatively new "groupware," which is now emerging in the commercial marketplace but is still strongly proprietary in nature and therefore not yet a good basis for interoperability or standardization. These systems have the advantage that they do not require the communicating parties to be linked simultaneously, thus overcoming the problems of geographic separation and time-zone differences. Anyone who has experienced either "telephone tag" or the trials of close collaboration with a colleague many time zones away can appreciate the utility of these asynchronous communication tools.

Recently, multimedia electronic messaging applications have been developed that allow complex documents, imagery, sound, and video information to be incorporated into messages to enrich the quality of the deferred communication. Technical problems still remain, however, in providing features such as automatic document format conversion between word processing systems. Even if technical solutions can be found, widespread deployment may be slow in coming since people are often reluctant to adopt new tools if the ones they are accustomed to seem to be serving them satisfactorily.

Video conferencing has been available for some years, but most such systems have required that users go to special conferencing centers to make use of cameras, monitors, and special communications equipment. Moreover, these systems worked in either broadcast mode or two-way, two-site interactive mode. *N*-way multiple-party video conferencing is more difficult to support. Nevertheless, a variety of services have been available commercially, using the telephone system for transmission.

Experimental and quasi-operational systems have been built to support video conferencing in a data network environment. The (Defense) Advanced Research Projects Agency, for instance, has been using an experimental packet-switched video-audio conferencing system on its wideband network for a number of years and recently transferred it to the Defense Information Systems Agency for more operational use. Recent experiments with packet-switched video and audio on high-performance work-

stations indicate that desktop conferencing is possible. This new technology, called multicast, has been developed and tested on the Internet. Multicast uses a TCP/IP packet-switched network to deliver copies of the video-audio packets to terminals on the network that have been temporarily designated to be part of the multicast conference through the use of special addresses. Once the multicast conference is over, the terminal resumes using its standard network address.

Interactive video may prove to be very important not only for conferencing, but also to support remote viewing of experimental procedures. Currently deployed research networks such as NSFnet do not have sufficient transmission and switching capacity to support very much video conferencing (although the more advanced National Research and Education Network (NREN) program contemplates such capabilities), and users may need special hardware to digitize and compress the video information before sending it on the data network. Workstation vendors are already developing such systems; relevant research and development are also under way at nearly all of the major workstation vendors.

More sophisticated and potentially more useful will be "shared workspaces," which would enable remote conversation with simultaneous joint viewing or other remote interactions. In the most general sense, shared electronic workspaces would mimic a complete physical research environment, with all of the data, software, and instrument control available to all parties within the context of a discussion. The coordinated data analysis workshops (CDAWs), discussed in Chapter 2, are an excellent example of a physically shared workspace. One can imagine a "virtual" CDAW in which many scientists participating from various geographic locations could interact with all sources, including all the participants and their data. Before such sophisticated computer-supported cooperative work environments can be realized, however, major technological advances must be made in areas ranging from wide-area network caching to multiuser synchronization.

More immediate modes of group interaction can be supported with existing telephone and audio conferencing technology and desktop video conferencing using specially equipped terminals with video cameras, often coupled with facsimile transmission or other deferred-communication tools to provide a context for real-time discussions. In addition, some tools for supporting group work and interaction—such as shared editors with synchronized displays linking multiple authors, or multiplayer games that allow each player to view the real-time movements of the others—are already available as research prototypes and introductory products from commercial vendors. However, such technologies are not necessarily broadly applicable to the data-intensive demands of scientific collaboratories.

Although most of the commercially promising computer-supported cooperative work technology does attempt to solve the problems associated with communication of voice, text, images, and data across networks (and in most cases joint authoring tools), it does not address remote control of instruments, remote collection of data, accessing of archived data, and resource discovery software. Computer-supported cooperative work technologies designed for commercial application are often implemented over private corporate networks operating in protocols and software environments that differ from those common in government-supported research networks. However, research that supports the development of commercial collaboration technology and the products that result will be of great interest and will likely aid the development of collaboration technology and collaboratories for science.

Achieving Transparency

It is highly desirable that the architecture of a collaboratory system be transparent, i.e., that it allow scientists to treat all the different databases, programs, instruments, and participants conceptually as being part of a single system by using a uniform set of commands accessible from their desktops or laboratory benches. To achieve this, the system must hide all its real-world variability internally. Prototype collaboratories discussed in the workshops (the Worm Community System and the Sondre Stromfjord testbed) demonstrate convincingly that it is technically feasible to implement and deploy such an architecture, at least on a relatively small scale.

Achieving truly transparent data access from federated databases will require research and development. Some technology for federating diverse data sources is available in research prototypes, such as the relatively small and focused Worm Community System, but providing these capabilities on the much larger scale that appears to be required for many scientists will require a major effort. The effort to develop Knowledge Robots (Knowbots™), which are essentially network-mobile intelligent agents, involves substantive research on distributed systems architecture and control, authentication, security, semantic representations, and a host of other problems in computer science (Kahn and Cerf, 1988).

Achieving database transparency is just one application of the much more general idea of cooperating intelligent agents working together over a computer-communications network. In the case of remotely controllable instruments, programmed intelligence is needed at the instrument site to accept control and prepare and deliver captured data. The more general case is that a set of programs distributed around the network must cooperate to achieve a particular objective. Distributed database systems, distributed processing systems, networked computing systems, and digital libraries are all instances in which cooperating intelligent agents could be and often are applied.

The excitement in the computing community over distributed, object-oriented, client-server systems results, in part, from recognition that these systems may help to break the constraints of time and distance in the conduct of scientific and other research (Wiederhold, 1992). Building collaboratories may be one of the most powerful ways of applying these new computing ideas in support of science.

ACKNOWLEDGING CONTEXT—SOCIAL AND INSTITUTIONAL CONSIDERATIONS

In addition to the technical expertise required to construct collaboratories, basic social and institutional factors must be examined and dealt with effectively to achieve successful scientific collaboration. Without sufficient attention to these potential constraints, in particular, the best collaborative information technology will have little positive impact on the working lives of scientists. Among the many issues to be addressed are the willingness and ability of individuals and institutions to engage in large-scale efforts of the kind needed to build useful collaboratories. Underlying these issues are questions about motivations for collaborating, the prospects for achieving a working partnership between computer scientists and other scientists, and the perceived trade-offs between the rewards and risks, financial and otherwise, of participating in collaboratories. These concerns must all be factored into the design and selection of collaboratory efforts.

Issues for Individual Scientists

To use and build collaboration tools and systems, individual scientists—both the users and developers of technology—must have the motivation to do so. Of particular concern is the perceived lack of opportunity for career advancement associated with electronic data sharing and collaboration. Further, collaboratories must be designed so that the rewards of electronic collaboration and data sharing outweigh the risks.

Current career paths may not gracefully accommodate scientists working in and building collaboratories. In the physical sciences and in computer science as well, fame accrues largely to the development of theories or ideas, and not to the development of tools or the gathering of data for others to use. Given long-standing traditions of individual achievement, and the more recent increase in competition for limited resources, collaboration is viewed warily by many scientists because neither collaboration in itself nor the facilitation of collaboration represents a direct means to gain acclaim, respect within the scientific community, or funding for research. Further, the objectives of individual

scientists may conflict with larger organizational and societal objectives, which may also conflict with each other. Thus, abstract arguments about such benefits of collaboration as better handling of complexity, scale, and/or interdisciplinary research problems may not lead easily to changes in individual behavior. Similarly, top-down mandates for collaboration are not likely to be productive as long as scientists continue to be rewarded almost exclusively for publishing results of self-initiated research, having their work cited in the literature, and otherwise becoming distinguished as individuals.

From a societal perspective, science advances through extensive, timely sharing of data (see Box 1.1). But to advance as individuals, scientists generally must use their own data to the fullest extent possible before sharing them with others. In addition, scientists need to be comfortable that data generated by others are of high quality and that any peculiarities associated with the data themselves or their collection are identified and understood. Given such constraints, it can be difficult for scientists to openly share data in recognition of a communal interest—that their community benefits from access to as much good data as possible.

Technology cannot solve these problems, but technology can be designed to mitigate them by making more data easier to use by more scientists. By facilitating the broader use of data collected by individual scientists, technology can enable research sponsors and the scientific community to leverage individual data collection investments. For this to happen, appropriate policies and rewards need to be established that support scientists who share data. For example, NASA-sponsored projects have well-defined "rules of the road" outlining the obligations of researchers with respect to use of spacecraft mission data (Appendix C), and NOAA has a similar set of guidelines for ocean-craft missions.

Another approach to rewarding and thus encouraging data sharing is to grant appropriate recognition for a contribution to an electronic archive or database, perhaps much in the way that a publication would be recognized. For sciences with more distributed data collection, such as genome research, publishers of journal literature may require that supporting data be deposited in the archives before articles referencing them can be published. This approach ties the additional reward of publication to data sharing, and many sequences are now submitted to databases very soon after discovery directly from many molecular biology departments. Nevertheless, the performance and management of electronic data archives must be trusted by scientists if such archives are to be used effectively. This implies a need for quality assurance and security (especially data integrity) mechanisms, some of which are procedural and some of which may involve the use of computer technology.[8] A third way to reward data sharing is for the scientific community and the funding agencies to explicitly support scientists who analyze or reanalyze existing data.

At the same time, information scientists or collaboratory builders have their own careers to manage, and they face, in their own context, reward and advancement issues parallel to those confronting other scientists. Traditionally, the prestige in science has gone not to the "technician" who develops a significant new tool but to the "scientist" who uses the tool to discover a significant new phenomenon (although sometimes these have been the same person). Furthermore, systems developers often have difficulty gaining tenure in academic computer science departments today; accordingly, they tend to gravitate toward industry.[9] Yet systems developers participating in building a collaboratory could find in the development of such technology a means for demonstrating to their own departments the scientific merits of these complex systems.

In some circumstances, collaboratories have the potential for creating a positive and reinforcing sense of community that increases as the scale, quality, and scope of the shared information in the collaboratory grows. Representing an effort perhaps analogous to constructing a major, landmark building or conducting a national project such as the space program of the 1960s, participating in the development and use of a collaboratory may confer a sense of shared purpose, teamwork, and community that can become self-sustaining.

Other second-order effects are also suggested by the experiences of scientific communities that already make extensive use of networking technology. The committee emphasizes that these are early effects because today's technology is primitive compared to what it envisions. Distributed groups

supported by technology can assemble expertise independent of the physical location of the scientists who possess that expertise. Network-based communication changes the character of informal exchange that scientists use to help them make sense of their work. Communicating the tips and techniques necessary to make experimental apparatus and data sets work as advertised need not depend on face-to-face exchange but can be shared electronically among broad communities of interest. Such sharing is quite common in electronic special-interest groups.

Collaboratories may lead to the creation of new electronic organizations just as the Arpanet and later the Internet led to the creation of large electronic groups. Many of the Internet groups are extraordinarily lively, with their own unique community identity and practices. However, with some notable exceptions, most of these do not produce any joint product of lasting economic or intellectual value. Their primary output is usually discussion.

In light of the importance of participants' motivations to achieving success, the first collaboratories should be developed with groups of scientists already predisposed to collaborate and to use collaboration technology. Biologists sequencing and mapping the genome of the nematode worm *Caenorhabditis elegans*, space physicists participating in CDAWs, and oceanographers involved in the Tropical Ocean-Global Atmosphere (TOGA) program illustrate what can be achieved when scientists themselves initiate collaboratory efforts.

Costs for Individual Scientists of Using Computer-based Collaboration Technology

Designers of collaboratories must recognize the costs and risks, as well as the benefits, associated with sharing data. Even for scientists who already see how computer-based collaboration technology can advance their work, the choice to use it (assuming it exists) is not a costless one. Based on the social history of computing to date, several kinds of costs need to be considered.

- *Incompatibility/critical mass of tasks and people.* Unless one's entire world is on-line, there will be inconveniences of switching from one medium to another. For example, if distributed project group members can share manuscript files that include data tables but not line drawings, or line drawings but not halftones, then at various points during the process of manuscript preparation some people will be denied access to the process.

- *Economic costs.* The history of organizational computing suggests that people continually underestimate the costs of operating, maintaining, and upgrading computer technology. These systems will require human support as well as capital and operating resources. If funds for direct personnel support are not forthcoming, they will show up as a tax on the time of scientific personnel. Doctoral students or postdoctoral researchers who are supposed to be doing science may end up spending a substantial fraction of their time doing technology support.

- *Dependence and vulnerability.* In addition to choosing and developing features that are responsive to scientists' personal and professional concerns about data sharing, collaboratory designers must also recognize that collaboratory performance overall will cause concern. The more one comes to depend on these technologies, the greater one's vulnerability when they break. At a minimum, technology failure engenders frustration. In some organizations today that are highly dependent on their internal networks, employees report that work absolutely stops when the network goes down. More seriously, technology failure may lead to irrevocable loss of data or work. Scientists will rightly have limited patience with experimental systems that may be unreliable; assurances as to system integrity and reliability will be needed. Ease of use and flexibility are other important systems dimensions.

Other costs relate to education, as discussed below. Even "easy" systems can be surprisingly hard to learn and use. These will not be easy systems. Nor will they be stable ones. Time invested in learning today's system features will not eliminate the need to spend time learning tomorrow's improvements.

Education and Training

Scientists and technologists must learn the necessary skills to use and to build collaboration tools and systems. Effective use of new technologies and facility in making a transition from old to new techniques typically require special training. This will certainly be the case for collaboratories. Both in research and commercial settings where collaboration technology has been introduced, training has been an important factor in the success of the implementations. Even experience with the Internet, which has not been particularly user-friendly, indicates that those who make the greatest use of the Internet have had to devote time and effort to learning how to do so. In the long term, collaboration technology and its use for remote interaction with data, software, instruments, and colleagues will become commonplace for research scientists, as has the use of personal computers and workstations. Until then, an investment of time, effort, and resources for training new students and practicing scientists should be considered a part of the process of launching collaboratories. Appropriate training might be provided in the form of predoctoral or postdoctoral fellowship programs, summer studies, visiting professorships, and research group exchanges.[10] The network itself may become the medium for delivering education and training to dispersed groups of scientists. Given the existing demands on scientists' time, the training burden must be minimized through the design of easy-to-use and easy-to-learn systems.

To develop genuinely useful technology, it will be necessary to train technical experts who can work closely with user/scientists. One approach is to create interdisciplinary programs of the kind started at Rice University (Box 5.2) that combine instruction in a specific scientific discipline with training and hands-on experience in computer and information science. The emergence of computational and mathematical biology as a subdiscipline provides additional insight into the need for special training. This example is explored more fully in Appendix D.

User-Developer Partnerships

A partnership between computer scientists and engineers, on the one hand, and scientists who recognize a need for better computing and communications capabilities, on the other, should provide intellectual and material benefits to both parties. The uneven support for computer-related infrastructure is a principal reason that scientists have often had to develop their own infrastructure, software tools, and applications. These systems and tools are often ingenious in their application of computing technology to science. However, designing and building tools may divert scientists from their primary area of research and may yield tools and systems that are less useful than the scientists would like.

Investigators at the frontiers of knowledge must be intimately involved in the design and definition of the tools they need to do their research. But if they can work in collaboration with skilled system builders, rather than act as their own programmers or programmer managers, they should be better served by the resulting tools. Furthermore, system builders may be better positioned to design and build generalizable tools that can subsequently be used by other scientists as well. Due to the specific needs of the scientists involved and the requirements of their science, the user-developer partnership will vary for different collaboratories. For some collaboratories it may be desirable for academic scientists to partner with industry to share resources and knowledge, and to facilitate technology transfer. For other collaboratories, computer engineers and software designers from the computer science research community may be essential. Although the mix of skills, talent, and training will thus need to be tailored

BOX 5.2 THE COMPUTATIONAL SCIENCE AND ENGINEERING GRADUATE DEGREE PROGRAM AT RICE UNIVERSITY

Rice University's Computational Science and Engineering (CSE) Graduate Degree Program is designed to provide interdisciplinary research and education in scientific computing.

The Master's Degree Program

The intent of the master's degree program is to graduate professional experts in scientific computing who will be able to work as technical specialists within an interdisciplinary research team. Degree candidates will be offered training in the use of state-of-the-art numerical methods, high-performance computer architectures, and software development tools for parallel and vector computers; application of these techniques to at least one scientific or engineering area; a curriculum consisting of topics from computer science, computational and applied mathematics, and a selected application area; and hands-on experience with leading-edge parallel supercomputers.

The Ph.D. Degree Program

The Ph.D. degree program offers the same interdisciplinary approach as the master's degree program but with greater specialization. An original thesis and, in addition, the completion of either an advanced schedule of courses or a computational project in an application area other than computer science or computational and applied mathematics, are required for the Ph.D. program.

Participating Departments

- Biochemistry and Cell Biology (expected)
- Computational and Applied Mathematics
- Electrical and Computer Engineering
- Chemical Engineering
- Computer Science
- Statistics (expected)

Admission

Students must be admitted into one of the academic departments listed above to be considered for participation in the CSE graduate degree program. The student participates as a graduate student within that department in every way except that the curriculum and examination requirements will be set by guidelines for the CSE graduate degree program.

SOURCE: Theresa Chatman, Center for Research on Parallel Computation, Rice University, Houston, Texas.

to meet the requirements of specific programs, the goal is to have scientists and technologists working together to develop the collaboratory infrastructure so that all may benefit equally. Variations in the approach to a partnership should be considered something positive, because adapting a collaboratory program to meet specific conditions and needs of a group of scientists will be essential to its success.

User-developer partnerships will present fundamental tensions that will have to be recognized and addressed. The primary interest of scientists will be the practice of their own science, whereas the primary interest of computer scientists will be in system design, prototype systems, and theoretical studies in support of system building. Without strong countervailing incentives, many scientists will be reluctant to invest too much of their time trying to use prototype systems that may be viewed as computer science experiments. Indeed, computer scientists will have to assure a satisfactory minimum level of performance for prototype systems that will affect other scientists' work and careers. Such issues have been faced and overcome before in the context of the Internet, in the development of expert systems for medicine such

as DENDRAL, in Stanford's SUMEX-AIM project,[11] and elsewhere, but reminders and reassurances on both sides of the partnership are likely to be necessary elements of the formation of a collaboratory program. All of this will require extra effort on the part of scientists and technologists. However, as demonstrated by pioneering efforts such as SUMEX-AIM or the more contemporary cases identified in the CSTB workshops, such extra effort can bring handsome rewards.

Issues for Individual Institutions

"Science regardless of distance" will not be cost-free to the institutions within which science is managed, funded, transmitted, and legitimized. Institutions must learn how to support their scientists who are working in collaboratories. It must also be recognized, however, that institutions may not, absent other changes, be willing or able to provide such support. The role of institutions will depend in part on the nature of available public infrastructure and the support they provide for such infrastructure.

In enabling the creation of new working relationships among users, collaboratories will offer both opportunities and problems for the institutions in which individual scientists are based. They are expected to result in new forms of organization that are dispersed in space and time and that may substantially affect the conduct and progress of science. Further, collaboratories offer the opportunity to produce electronic products and services with economic and intellectual value. Issues related to intellectual property rights will undoubtedly arise as collaboratories become widely used in academic and industrial laboratory settings. It is impossible to say how the legal and economic issues associated with the work done through collaboratories will evolve, but it is clear that use and development of collaboratories must take these issues into account.

Providing Local Infrastructure Support

Today's division of responsibility and labor for supporting computing and communications infrastructure for science grew out of yesterday's technology and is demonstrably inadequate as technology rapidly changes. This inadequacy is highlighted by the question, Who pays for what? For example, research grants may pay for workstations in a laboratory but not for connections to the campus network. Campus communications organizations may pay for wiring between, but not within, buildings. The federal effort has built a national backbone network for research, but the NSFNET backbone per se excludes local network domains within universities. A university library may pay for journal subscriptions but not for searches of on-line literature databases. In oceanography research grants may pay for electronic mail services, but in space physics they may not. For an infrastructure to function, all the components must be adequately funded and supported; if any are missing, the entire system is weakened and can fail. Although the development of a national information infrastructure, for researchers and educators (e.g., through the NREN) and for the economy as a whole, is attracting attention to nationwide connectivity, it is essential to provide also for the local components—the access points, the support and maintenance capabilities and costs, local user training, and provision at each site of reference and tutorial documentation.

Having a particular technology does not guarantee that it can be used effectively. Providing support extends beyond building or purchasing networks and tools to ensuring the availability of people skilled in organizing and providing support services. Such services may range from maintaining help desks and hot lines to providing information management support that will save researchers time and effort by helping them to set up workstations, connect to the network, install the software, use the features, update the databases, and so on. For information infrastructure to become widespread and institutionalized, there must be adequate and standard funding for support and maintenance.

Discussions during the three workshops conducted for this project underscored the value of technicians and other paid professionals and paraprofessionals who provide essential technical support services but who, under current conditions and despite the best of qualifications or contributions, have a secondary status in the projects they support and often in the academic career structure of the department in which they work.[12] Workshop participants acknowledged a lack of financial incentives and career advancement opportunities as an impediment to attracting and keeping talented support personnel. This is an issue for the management of institutions, and also for the proposed collaboratories, which would involve both the conventional range of support personnel and, also complementing the domain scientists, computer scientists and engineers who, as professionals in their own right, would expect to function on a par with their scientist partners.

Managing the Results of Increased Interaction

Using collaboration technology, individual scientists will have the capability to form new working relationships independent of their physical institutional homes. While these relationships may energize scientists and lead to new discoveries, they may also complicate the role of managers in scientists' home organizations who are charged with keeping track of scientific manpower and resources. Issues of control and accountability may arise for organizations whose members can form new collaborations or access new resources at will (Box 5.3). Although institutions may already have mechanisms for overseeing or engaging in projects involving many institutions, these mechanisms may be too cumbersome for the kinds of fluid, evolving relationships that will be possible in collaboratories.

BOX 5.3 FACTORING THE NINTH FERMAT

"Two mathematicians employed by Bell Communications Research (Bellcore) and Digital Equipment Corporation used electronic mail to recruit [computing resources from] several hundred researchers from companies, universities, and government laboratories around the world. They asked them to work on solving a large and important mathematical problem, one with practical implications for cryptography. Researchers who volunteered to help were sent a piece of the problem and returned their solutions by electronic mail. All of the partial solutions were then used to construct the final solution. The electronic message announcing the final results contained a charming admission: the two mathematicians who organized the work and constructed the final solution from the pieces returned to them did not even know the names of all of the people who helped them:

> We'd like to thank everyone who contributed computing cycles to this project, but I can't: we only have records of the person at each site who installed and managed the code. If you helped us, we'd be delighted to hear from you; please send us your name as you would like it to appear in the final version of the paper. (Manasse, 1990)

[This case highlights some of the limitations in conventional thinking about organization and management that may become apparent when networked organizations become more common.] . . . Typically managers influence their subordinates in large measure by allocating resources to their projects and allocating credit (or blame) to their accomplishments. How will the manager's role in resource allocation change when people can reach out across the network and directly solicit resources from others to help them with their work? How will the manager's role in allocating credit or blame change when managers do not know, and perhaps cannot know, who contributed in what ways to accomplishments?"

SOURCE: Sproull and Kiesler (1991), pp. 160-161.

Collaboratories, particularly by means of electronic-mail discussion groups, can foster the discovery of potential cooperative or collaborative partners. As a consequence of day-to-day interaction in an electronic environment, participants who might not normally meet face to face may learn of mutual interests that may lead to direct contact, laying the groundwork for more extensive cooperation or collaboration. Anecdotal evidence supports this view. In the early years of the Arpanet project sponsored by (D)ARPA, it was thought that the use of networked electronic mail would reduce the need for travel, but it later became apparent that travel budgets had actually increased substantially. One reason was that electronic mail allowed more people to interact directly and also allowed geographically larger projects to be managed; such projects became cost-effective as well as feasible because of facilitation via electronic mail. When the participants did get together, there were more of them and they typically traveled longer distances.

ISSUES IN FUNDING FOR COLLABORATORY INFRASTRUCTURE

The major High Performance Computing and Communications (HPCC) program recently initiated by the federal government provides a favorable context in which to consider a serious effort to develop collaboratories. Aimed at attacking many of the "grand challenges" of science (Federal Coordinating Council for Science, Engineering, and Technology, 1992), the program promotes the concept that HPCC technologies are fundamental to progress in many areas of science, especially advances in computational science, and articulates the vision of a national scientific information infrastructure. Collaboratories complement and build upon HPCC activities and technologies. Specifically, a collaboratory program, while distinct from the HPCC initiative, would drive research in advanced software technology, tools, and especially scientific applications of the NSFNET, NSInet, and other constituents of the Internet engendered under the NREN program. Even in rudimentary forms, collaboratories make the existing nationwide research networks more useful. Collaboratories can thus leverage investments in HPCC technologies and applications, and they can assist in the establishment of a national information infrastructure for science.

Scientists' needs for computing and communications support have not been met in part because of an absence of mechanisms for funding information infrastructure development.[13] While funding for infrastructure has been inadequate at the program level in most funding agencies, the greatest obstacle is structural. Scientific funding agencies are organized along discipline or mission lines. Although units within these agencies recognize the need to support major instruments and facilities unique to a field or mission, no single unit is responsible for funding general computing and communications infrastructure within a given field. Only NSF has a mission that supports basic science across all fields, but its resources for infrastructure are limited.

Precisely because infrastructure is useful to everyone, it seems to be the responsibility of no one in the research funding structure. A consequence is that if program officers set aside resources for infrastructure, they do so at the apparent expense of research in their discipline. Faced with this dilemma, an individual program officer finds it almost impossible to make the decision in favor of infrastructure—even if that decision is in the long-term interest of the field. This structural barrier to providing adequate ongoing funding for infrastructure must be overcome if the country is to create, maintain, and benefit from the premier information infrastructure that many now want to build.[14]

Funding a collaboratory program does not have to imply taking money away from individual research grants. First, in recognition of the problems that many scientists have in collaborating under any circumstances and the added problems anticipated from broadening collaboration to include computer scientists and engineers along with domain scientists, the committee envisions that collaboratory projects will be selected from the bottom up. That is, they will be launched in response to inquiries and efforts by groups of scientists who recognize a need to collaborate and who manifest an interest in applying more and better information technology. This pattern has been successful, for example, in biology research

projects associated with Los Alamos and Brookhaven National Laboratories. Second, since collaboration and the use of information technology are already elements of existing or planned projects (e.g., the HPCC program, the Sondre Stromfjord Observatory, the Solar-Terrestrial Energy Program, the World Ocean Circulation Experiment, and CDAWs), collaboratories could be incorporated into such projects more economically than starting ab initio. Third, economies from easier data sharing would free up resources otherwise spent on duplicative efforts to gather or store data.

The committee is sensitive to a primary concern of many scientists that already-tight money for research not be diverted. However, it wishes to emphasize that collaboratories are most likely to succeed in those areas where research is already being transformed into an activity that cannot be done without the cooperative application of information technology. In those areas, investment in collaboratory projects, through conscious development and use of technology, should have as a payoff the more efficient and effective use of research resources.

MAKING A START

Information technology to support collaboration will not impel collaboration, but it can enable it. Scientists will use that technology if they believe it will advance their own work, and if they perceive the benefits to their work as outweighing the costs of using the technology. The likeliest cases of net benefit are those in which scientists who have already chosen to collaborate can be supplied with technology that makes it easier—in terms of cost in time and effort—for them to do what they are already doing. Moreover, some scientists will recognize that collaboratories, in some instances, can enable better research and greater exploration of specific topics.

Efforts to date have hardly scratched the surface of the potential that collaboratories offer. The most significant results will be achieved with sustained and focused efforts to develop valuable shared databases and digital libraries of specific scientific content; to develop collaboration tools for particular scientific purposes; and to put into operation institutionalized pilot collaboratory programs, building on rapidly evolving technology trends. The technology base for such an effort is ready, the programmatic environment is especially auspicious, the computer and communications research community is particularly interested in making this a high-priority item, and the national interest will be well served by developing leading-edge information infrastructure that can serve not only the scientific community but, eventually, the business sector as well. We now have sufficient knowledge to substantially improve the functionality of collaboration technology, if only the scale of the efforts can be made adequate to the task of demonstrating their effective utility.

NOTES

1. Archie—Index of FTP archies and file fetcher "Archie.doc", on-line @ cs.washington.edu. Author(s): Archie—Alan Emtage, Peter Deutsch, Bill Whelan @ cs.mcgil.ca; Prospero—Clifford Neuman @ isi.edu; Archie client—Brendan Kehoe @ cygnus.com.

2. Gopher—Distributed document delivery service. "Gopher/doc/client.doc & server.doc", on-line @ boombox.micro.umn.edu. Author: The Internet Gopher Team, University of Minnesota, Minneapolis.

3. World-Wide Web—Distributed hypertext information server. "The World Wide Web", on-line @ info.cern.ch. Author: Tim Berners-Lee, World Wide Web Project, CERN, 1211 Geneva, Switzerland.

4. Wide Area Information Service—Distributed text search and retrieval service. "WAIS Overview", on-line @ think.com. Authors: Harry Morris, Brewster Kahle, Jonathan Goldman, Thinking Machines Corp.

5. For example, the Worm Community System uses a representation called an information space, in which each item of data is transformed into a uniform object called a unit of information and there is a standard representation for describing links between information units.

6. The calling sequence of a program is essentially the information that must be passed to the program for it to be run. For example, the calling sequence of a subroutine must specify its arguments and parameters; the calling sequence for a program must specify its location (e.g., its home directory). The calling sequence can be specified in many ways and is a matter of convention dictated by the structure of the computing environment in which the program will run.

7. One such standard, ASN.1 (for Abstract Syntax Notation version one), was developed by the Organization for International Standardization (ISO) for the syntactic exchange of data. At the National Library of Medicine, ASN.1 is used in the National Center for Biotechnology Information toolkit to support semantics for several common biology types and is being adopted in a number of analysis software packages in molecular biology as a common data interchange format (Ostell, 1992). In other contexts, notably those involving high-performance computing, ASN.1 does not offer satisfactory performance.

8. A range of quality control processes exists for different purposes, from moderating (checking for topic) conference proceedings or newsletters to editing (checking for accuracy) journals or books. For journal literature, the peer review process ensures quality. For data generated by instruments operated by a single investigator or team, as is commonly the case in space physics and oceanography, the investigator is typically responsible for quality control and thus performs both data contribution and checking. In genome projects, with more distributed data collection than in oceanography and space physics projects, contributions to the archives have traditionally only been moderated and not edited, although large databases now have a curator whose responsibility it is to review submissions and maintain quality control for the information that will ultimately reside in the database.

9. A forthcoming Computer Science and Telecommunications Board report will examine this issue in the context of academic careers for experimental computer scientists.

10. See Appendix D, a reprint of Appendix 3 of the final report of an NSF-sponsored workshop, Training Computational and Mathematical Biologists, held at the Banbury Center of the Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, December 9-11, 1990.

11. The SUMEX-AIM project, begun in 1974, brought together scientists and technologists to facilitate research on the applications of artificial intelligence to medical (AIM) research. The project used the Arpanet and Tymnet to link together the Stanford University Medical Experimental computer (SUMEX) and a group of researchers from around the country. Although the computing resources available were primitive by today's standards, the project used electronic mail and an electronic bulletin board. The result of this collaboration was an encyclopedia of artificial intelligence tools, the *AI Handbook*. According to one description of the project,

> [S]uch a resource offers scientists both a significant economic advantage in sharing expensive instrumentation and a greater opportunity to share ideas about their research. This is especially timely in computer science, a field whose intellectual and technological complexity tends to nurture relatively its own line of investigation with limited convergence on working programs available from others. The complexity of these programs makes it difficult for one worker to understand and criticize the constructions of others, unless he has direct access to the running programs. In practice, substantial effort is needed to make programs written on one machine available on others, even if they are, in principle, written in compatible languages. In this respect, computer applications have demonstrated less mutual incremental progress from diverse sources than is typical of other sciences. The SUMEX-AIM project seeks to reduce these barriers to scientific cooperation in the field of artificial intelligence applied to health research. (Lederberg, 1978)

12. Some support services are as informal as the graduate student who figures out and shares a trick for working with a particular program, system, or database.

13. Although information infrastructure is now receiving increased attention as a matter of public policy and private enterprise, current efforts provide scientists with limited access to specialized network-based computing tools. The NSF supports the supercomputing centers, and to a lesser extent the science and technology centers, to assist scientists in using specialized, state-of-the-art hardware and software. NSF has also sponsored the development of the NSFNET backbone network and tributary, intermediate-level networks. It has maintained a modest program to assist institutions in acquiring the capital equipment needed to link to the Internet. Other government agencies such as the DOE, NASA, (D)ARPA, and NOAA provide assistance to their scientific communities, but overall access to network-based computing tools is still limited.

14. The anticipated interagency task force on information infrastructure might address this issue, although its focus is expected to be meeting the needs of the general public (industry, nonprofit organizations, and individuals).

6

Providing for a National Collaboratory Program

Computing and communications tools are already being adopted by a wide variety of scientists to automate data retrieval and analysis, to handle the growing scale and complexity of scientific problems, and to facilitate sharing of data and insights with colleagues. Thus a foundation has been laid for more systematic and more integrated uses of computer-based tools in scientific applications where collaboration is necessary or desirable. Achieving this next level in the use of technology, referred to as the collaboratory concept, will involve both technical and social challenges. Computer scientists and engineers will have to join with scientists in other fields in partnerships focused on improving collaboration in specific scientific arenas. Effective partnerships promise to yield both gains in the productivity of collaborating scientists and gains in scientific achievement in areas where the technology makes possible the exploration of new kinds of questions and the use of new methodologies for scientific research. At a time when the cost of scientific research is expanding while research support is tightening, such gains take on a special appeal. The committee believes that the time is right for a focused initiative to pursue scientific collaboratory projects and develop associated technologies.

As discussed in Chapters 2 through 4, several prototype collaboratories have been developed for limited applications. These prototypes have demonstrated the usefulness and viability of the collaboratory concept as a means of facilitating scientific research. Based on its findings, the committee recommends an initiative to:

Establish a research program without delay to further knowledge of how to build, operate, and use collaboratories in the support of science. This program should have two major components of equal importance:

- **A *research component* dedicated to developing and integrating the software and hardware needed to build and apply collaboratories.**

- **An *education component* dedicated to educating and training the people needed to build and use collaboratories.**

The overarching goal of the program is to aid science and scientists through the construction and operation of working collaboratories. To achieve this goal the committee further recommends that the program:

Establish several collaboratory testbeds, funded at a level of $6 million per year each over a period of 5 years each.

The committee has learned from the workshops held in oceanography, space physics, and genome mapping and sequencing that these particular fields of science can benefit from the use of collaboratories,

and it is clear that opportunities exist for their use in other disciplines such as seismology and neuroscience (Institute of Medicine, 1991). Concurrent construction of collaboratory testbeds (Box 6.1) will encourage synergistic technology development and permit several sciences to explore the technology in the short term. At the same time, generic collaboratory technology will benefit from being developed for and adapted to several disciplines.

Multiple testbeds tailored to the needs of particular disciplines are recommended to investigate thoroughly the use of collaboratories to discover, teach, and transfer scientific knowledge. Testbeds are the only effective way to explore the multifaceted nature of these tool-oriented computing and communications systems. The committee believes that these testbeds can play a major role in demonstrating how science will be done in the 21st century and how a national program for information infrastructure for research can be implemented.

A collaboratory testbed program has the potential to support science in at least four different ways: (1) by giving scientists tools to do more and better science; (2) by giving teachers tools that they and their students can use to experiment, explore, and collaborate; (3) by involving industry in collaboratory development, thus giving scientists a means to transfer technology from the laboratory to the business sector, which can then make collaboratory technologies and services commercially available; and (4) by providing opportunities to understand better the social and organizational dynamics of scientific research conducted using collaboratories. Such research can be used to refine and improve collaboratories, making them more useful to science and more easily used by scientists. An important consequence of such a program will be the development of the human resources needed to support the building and operation of scientific collaboratories.

A program length of 5 years is recommended due to the nature of the project. Building collaboratories will require that existing technology be adapted and integrated in new ways, and that new technology be developed. A significant period is needed in which to develop and apply the technology and to refine it based on experience.

It is expected that each testbed will involve senior natural scientists, senior computer scientists, senior social scientists, and a variety of junior-level people. The committee estimates that about 50 full-time-equivalent (FTE) workers will be needed per testbed to create a critical mass of expertise. Fifty FTEs will cost about $5 million per year, based on a mean cost per FTE of $100,000, including overhead. Further, 50 FTEs will need about $1 million for computer equipment and networking facilities per year averaged over the 5-year period of the program. Hence the total annual cost is estimated at $6 million per testbed. The relatively high equipment capitalization costs are a consequence of intensive use of computing technology for instrument control, large-scale databases, and scientific visualization tools associated with the kinds of oceanographic, meteorological, and genome mapping collaboratories considered by the committee. The $6 million estimate assumes that testbed collaboratories will make use of existing scientific instruments and equipment. The committee notes that the NSF supercomputer centers, the NSF science and technology centers, and the NIH and DOE Human Genome Project require similar levels of support to accomplish comparably ambitious objectives.[1]

A program involving three testbeds implies a total funding level of $90 million to $100 million. The committee believes that this level would enable the program to achieve critical mass. However, it is recognized that this is a time of tight resources. Although shrinking the scale of the overall program somewhat may be a necessity, the committee emphasizes the importance of providing sufficient resources per testbed to achieve sufficient scale (see Box 6.1). Prototype collaboratories and technologies examined by the committee have been developed and implemented on a small scale; the challenges of developing and implementing collaboration technologies on a larger, effectively national, scale are substantial. Consequently, the committee formulated its recommendation to include a minimal number of testbeds while providing for a level of resources per testbed that, based on the experience of other projects supporting collaborative research, appears necessary to achieve success. Given the importance of scale, undertaking fewer testbeds at the same time may be one approach to stretching resources.

BOX 6.1 COLLABORATORY TESTBEDS—IMPACT OF SCALE

The purpose of a collaboratory testbed is to build and propagate a complete collaboration system to a specialized community of scientists. The goal is to run a large-scale experiment in both the technology of building a collaboratory and the sociology of using it, in order to infer what some of the components of an effective national research information infrastructure should be. Funding must accordingly be available for a period sufficient to run a large-scale experiment.

As shown by the limited scale of staffing for current collaboratory "testbeds,"* these testbeds are really prototypes. They include a small sample of users, implement a small sample of a technology, bring the users up on the technology with minimal support, and track a sample of the usage. For example, the Worm Community System is currently only a model implementation and will have to be expanded considerably before it can function as a complete information infrastructure. In particular, the sets of data libraries and analysis programs must be expanded, a true distributed system across platforms and networks must be developed, and a case-hardened implementation must be evolved before the technology can be used as a springboard for a generic infrastructure. The expansion factor from prototype to testbed is partially due to increased functionality but primarily due to increased service, in providing more complete coverage and availability. In communications system development, it is common for the functionality to require only 10 percent of the effort, while universality and maintainability require the other 90 percent. So it is not surprising that the existing collaboratory "testbeds" are experiments in feasibility, rather than complete testbeds.

Expanding a prototype into a testbed involves supporting a much wider range of data types, data sources, hardware platforms, user interfaces, analysis programs, explanatory documents, and technical support. It would also involve defining the interfaces and providing the toolkits to permit many scientists to add their own wide range of data, programs, and interactions.

To see what might be necessary for a complete collaboratory testbed, it is instructive to examine centers for production systems. For example, the Genome Data Base, the central archive for the genetic data from the Human Genome Project, is a production project that provides reliable retrieval and periodic updates to a broad user community with significant user support, i.e., continuously running, daily updates of information contributed by several thousand users. Communications infrastructure projects, such as Arpanet and Internet, which are the direct predecessors of collaboratories, began as research projects at a few sites, grew into research prototypes at vertically integrated, quasi-academic organizations, and then became institutionalized as production developments at commercial corporations or, in the case of Arpanet, in the U.S. defense command and control system.

The development of collaboratories will likely move through this same path; the research projects are just now beginning with prototype collaboratory efforts such as the Worm Community System, the Sondre Stromfjord collaboratory, and highly collaborative research efforts such as the Tropical Ocean-Global Atmosphere and World Ocean Circulation Experiment programs. Prototype and testbed efforts point the way toward an institutional stage. Collaboratory development needs to become institutionalized, with resources for it built into national science funding policies.

---

*Current collaboratory "testbeds" supported by the NSF's program in collaboration technology are funded at about $0.6 million per year. For example, the Worm Community System currently receives about $400,000 per year, evenly split between systems building research of the core infrastructure and computer science research into longer-term technology. The Sondre Stromfjord facility in space physics currently receives about $800,000 per year, fairly evenly split between computer scientists developing the core infrastructure and domain scientists programming the specialized instruments. The corresponding staffing levels are thus about 4 full-time-equivalent employees for the Worm Community System and about 8 for the Sondre Stromfjord facility.

Consistent with the federal government's tradition of supporting basic research, the collaboratory program is envisioned as a federal program, building on related research efforts at NSF and various mission agencies such as NASA, DOE, and DOD. However, consonant with the "partnership" vision, a collaboratory program should be designed to draw on efforts and support from other sectors. There are a variety of options for industrial participation, for example, since industry not only has begun to develop commercial collaboration technology, but also employs people engaged in collaborative research (sometimes in collaboration with academic researchers). Universities not only house and benefit from scientific centers, but also employ scientists who collaborate with distant colleagues. Scientific organizations support and promote a variety of research programs, some of which (especially those viewed as "big science") require collaboration and several of which are under budget pressures. The full range of organizations that can benefit from collaboratories should be recognized and leveraged; a collaboratory program should be designed to encompass contributions of personnel, equipment, research facilities, and other resources from industry, academia, federal laboratories, and scientific organizations. Those contributions could both diminish and leverage the federal contribution.

As a part of the technology component of the program, the committee further recommends that the program:

Provide for two national demonstrations of each collaboratory testbed.

The national demonstration would provide a showcase of the program to the larger scientific and technical community. Such a demonstration could be held in conjunction with the meeting of a large scientific organization, such as the American Association for the Advancement of Science, to reach a wide-ranging audience or could be presented at discipline-specific conferences to promote and focus discussion within the community using the collaboratory. A national demonstration serves at least three purposes. First, it motivates participants in the testbed program to meet deadlines and drive the technology development for working collaboratories. Second, it educates the scientific community about how collaboratories can be used to advance science. Third, it provides scientists and technologists further opportunities to interact and collaborate. An initial national demonstration of each testbed could take place about half way though the proposed 5-year project cycle. A second demonstration could take place at the end of the cycle.

The cost of this feature of the program would be about $2 million per demonstration per testbed, based on the committee's estimates of the costs for other similar national demonstrations. A major value of such demonstrations is that the results are not ephemeral, but are integrated into and become a part of each demonstrated testbed.

As part of the education component of the program, the committee also recommends that plans be made to:

Initiate multiple and complementary activities to develop the human resources needed to carry out the collaboratory program, including, but not limited to:

A summer fellowship program to provide hands-on training for scientists and technologists in the use and development of collaboratory technologies in the conduct of science.

The summer fellowship program is aimed at increasing the community of scientists and technologists experienced in the development, implementation, and application of collaboratories and collaboratory technology. Fellowships could be sponsored by scientific societies, such as the American Association for the Advancement of Science, and the government in conjunction with academic institutions. Fellowships also provide a forum for interdisciplinary training of scientists and technologists. Such training is crucial to the success of the collaboratory program.

The cost of the fellowship component of the collaboratory program is estimated at $1 million per year, an amount that would support 50 summer fellows per year at $20,000 each.

Regularly scheduled national symposia for testbed principal investigators, research staff, and graduate students, providing opportunities to share information, findings, and conclusions regarding the technical aspects of building, operating, and using collaboratories.

The objective of this recommendation is to foster the creation of a community of builders and users of collaboratory technology. It is envisioned that principal investigators, research staff, and graduate students working on the testbeds would attend these meetings to share experiences and the results of their work.

In conclusion, the committee believes that the program outlined by these recommendations is the appropriate level of effort that should be undertaken. Although it would be possible to carry out only the research component of the proposed program, the committee believes that the education component is a critical aspect of the effort. Without the education component, it is much less likely that a skilled and growing community of collaboratory users and developers will emerge. The two components taken together as a program constitute a strategic effort to improve the basic infrastructure supporting the conduct of computationally intensive science in the United States.

NOTE

1. For example, the personnel budget of the National Center for Supercomputing Applications (NCSA) at the University of Illinois is about $6 million per year, of which about $5 million supports about 110 full-time senior people (personal communication with NCSA staff, January 15, 1993). Of these, about 50 full-time employees are involved in software development or scientific applications; the rest provide infrastructure support, consulting, documentation, and other user services. The nonsalary component of NCSA's funding is devoted primarily to equipment purchase or lease and maintenance, including staff, infrastructure, and support costs. The major funding for the Human Genome Project is a grant from NIH and DOE for $6 million per year, which provides support for about 50 full-time senior people, who are primarily at the main site at Johns Hopkins University, plus coordination of the many other consultants and contractors (personal communication with John Wooley, DOE).

References

Adams, M.D., J.M. Kelly, J.D. Gocayne, M. Dubnich, M.H. Polymeropoulos, H. Xiao, C.R. Merrill, A. Wu, and R.F. Moreno. 1991. "Complementary DNA Sequencing: Expressed Sequence Tags and the Human Genome Project," *Science* 252:1651-1656.

Altschul, S.F., W. Gish, W. Miller, E.W. Meyers, and D.J. Lipman. 1990. "Basic Local Alignment Search Tool," *Journal of Molecular Biology* 215(3):403-410.

Baker, D.N., R.D. Zwickl, and J.L. Green. 1984. "NASA Data Systems Users: Recommendations for Improved Scientific Interactions," *EOS* 65:46.

Briscoe, M.G., and D.E. Frye. 1987. "Motivations and Methods for Ocean Data Telemetry," *MTS Journal* 21(2):42-57.

Brooks, David A., and Melbourne G. Briscoe. 1991. "Tests of Long-Range Ocean Data Telemetry Using Frequency-Agile HF Packet-Switching Protocols," *J. Atmos. Oceanic Technol.* 8(6):856-858.

Burks, C., J.W. Fickett, W.B. Goad, M. Kanehisa, F.I. Lewitter, W.P. Rindone, C.D. Swindell, C-S. Tung, and H.S. Bilofsky. 1985. "The GenBank Nucleic-Acid Sequence Database," *Computer Applications in the Biosciences* 1(4):225-234.

Cantor, Charles, and Sylvia Spengler. 1992. "Primer on Molecular Genetics," Appendix A (pp. 192-218) as revised and expanded by Denise Casey in *Human Genome: 1991-92 Program Report*, DOE/ER-0544P, U.S. Department of Energy, Washington, D.C., June.

Carhart, R.E., S.M. Johnson, D.H. Smith, B.G. Buchanan, R.G. Dromey, and J. Lederberg. 1975. "Networking and a Collaborative Research Community: A Case Study Using the DENDRAL programs," pp. 192-217 in P. Lykos, Ed., *Computer Networking and Chemistry*, American Chemical Society Symposium Series No. 19, ACS, Washington, D.C.

Cinkosky, J., J.W. Fickett, P. Gilna, and C. Burks. 1991. "Electronic Data Publishing and GenBank," *Science* 252:1273-1277.

Dickey, Thomas D., 1991. "The Emergence of Concurrent High-resolution Physical and Bio-optical Measurements in the Upper Ocean and Their Applications," *Rev. Geophys.* 29:383-413.

Dickey, Thomas D., Timothy C. Granata, and Isabelle Taupier-Letage. 1992. "Automated in Situ Observations of Upper Ocean Biogeochemistry, Bio-optics, and Physics and Their Potential Use for Global Studies," Ocean Climate Data Workshop, February 18-21, Goddard Space Flight Center, Greenbelt, Md.

Dickey, Thomas D., R.H. Douglass, D. Manov, D. Bogucki, P.C. Walker, and P. Petrelis. 1993. "An Experiment in Two-Way Communication with a Multi-variable Moored System in Coastal Waters," *J. Atmos. Oceanic Technol.*, in press.

Federal Coordinating Council for Science, Engineering, and Technology (FCCSET), Committee on Physical, Mathematical, and Engineering Sciences. 1992. *Grand Challenges 1993: High Performance Computing and Communications*, Office of Science and Technology Policy, Washington, D.C.

Fields, C. 1992. "Data Exchange and Inter-database Communication in Genome Projects," *Trends in Biotechnology* 10(Jan./Feb.):58-61.

Gilbert, Walter. 1991. "Towards a Paradigm Shift in Biology," *Nature* 349(10 January):99.

Gish, Warren, Jonathan Kans, James Ostell, Gregory Schuler, and Karl Sirotkin. 1992. *Programmer's Reference: NCBI Software Development Kit*, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Md.

Green, J.L., J.H. Waite, J.F.E. Johnson, J.R. Doupnik, and R. Heelis. 1983. *Proceedings of the 1st Space Plasma Computer Analysis Network (SCAN) Workshop: Space Plasma Computer Networks*, NASA-TM-82514, Marshall Space Flight Center, Huntsville, Ala.

Green, J.L., and J.H. King. 1986. "Behind the Scenes During a Comet Encounter," *EOS* 67:105.

Greenstadt, E.W., and J.L. Green. 1981. "Data Systems Users Working Group," *EOS* 62:59.

Hardy, Darren R., and Michael F. Schwartz. 1993. "Essence: A Resource Discovery System Based on Semantic File Indexing," 1992 Winter USENIX, January 25-29, 1993, San Diego, Calif.

Institute of Medicine. 1991. *Mapping the Brain and Its Functions: Integrating Enabling Technologies into Neuroscience Research*, National Academy Press, Washington D.C.

Kahn, Robert E., and Vinton G. Cerf. 1988. *The Digital Library Project, Volume 1: the World of Knowbots*, Corporation for National Research Initiatives, Reston, Va.

Lander, Eric S., Robert Langridge, and Damian M. Saccocio. 1991. "Mapping and Interpreting Biological Information," *Commun. ACM* 34(11):33-39.

Lederberg, Joshua. 1978. "Digital Communications of the Conduct of Science: The New Literacy," *Proceedings of the IEEE* 66(11):1314-1319.

Lederberg, Joshua (compiler). 1990. *The Excitement and Fascination of Science: Reflections by Eminent Scientists*, Vol. 3, Part 1, Annual Reviews Inc., Palo Alto, Calif.

Manasse, M.S. 1990. "Complete factorization of the ninth Fermat number." Electronic message, June 15.

National Research Council (NRC). 1990a. *Interdisciplinary Research: Promoting Collaboration Between the Life Sciences and Medicine and the Physical Sciences and Engineering*. National Academy Press, Washington, D.C.

National Research Council (NRC). 1990b. *The Ocean's Role in Global Change: The Contemporary System*, National Academy Press, Washington, D.C.

National Research Council (NRC). 1990c. *TOGA: A Review of Progress and Future Opportunities*, National Academy Press, Washington, D.C.

National Research Council (NRC). 1990d. *The U.S. Global Change Research Program: An Assessment of the FY 1991 Plans*. National Academy Press, Washington, D.C.

Office of Science and Technology Policy (OSTP). 1992. *The National Research and Education Network Program: A Report to Congress*, OSTP, Washington, D.C.

Ostell, J. 1992. *Entrez: Sequences User's Guide*, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, April, Bethesda, Md.

Rees, David, Israel Perla, Nigel Meredith, James Green, and Nick Van der Heijden. 1986. "Networking Ground-based Images of Halley During the Giotto Encounter," *EOS* 67:1385.

Sanderson, T. 1990. "Networks and the Space Scientist," pp. 356-361 in *Computer Networks and the ISDN Systems*, Elsevier, North Holland, Amsterdam.

Schatz, Bruce R. 1987. "Telesophy: A System for Manipulating the Knowledge of a Community," Proceedings of the IEEE Globecom '87, Tokyo, November, pp. 1181-1186.

Schatz, Bruce R. 1989. "Searching in a Hyperlibrary," Proceedings of the 5th IEEE International Conference on Data Engineering, Los Angeles, February, pp. 188-197.

Schatz, Bruce R. 1991-1992. "Building an Electronic Community System," *Journal of Management Information Systems* 8(Winter):87-107.

Sproull, Lee, and Sara Kiesler. 1991. *Connections: New Ways of Working in the Networked Organization*, MIT Press, Cambridge, Mass.

Steele, Guy. 1984. *Common LISP, the Language*, Digital Press, Bedford, Mass.

Thomas, V.L., and J.L. Green. 1987. "Space Physics Analysis Network: A Quick-Reaction Capability," *Information Systems Newsletter* 13:30.

Thomas, V.L., J.L. Green, and B. McLendon. 1987. "SPAN Provides Electronic Transmission of Supernova Data," *NSSDC News* 3(2):1.

Thomas, V.L., and J.L. Green. 1988a. *SPAN Justifications*, NSSDC Technical Report 88-22, National Space Science Data Center, Goddard Space Flight Center, Greenbelt, Md.

Thomas, V.L., and J.L. Green. 1988b. "SPANning the Globe," *Information Systems Newsletter* 15(December):21.

Towards a National Collaboratory. 1989. Unpublished report of an invitational workshop held at the Rockefeller University, March 17-18, 1989 (Joshua Lederberg and Keith Uncapher, co-chairs).

Webster, Peter J., and Roger Lukas. 1992. "TOGA COARE: The Coupled Ocean-Atmosphere Response Experiment," *Bull. Am. Meteorol. Soc.* 73(9):1377-1416.

Wiederhold, Gio. 1992. "Mediators in the Architecture of Future Information Systems," *Computer*, March, pp. 38-49.

Winterhalter, S. 1986. "Technology Development Activities Support Voyager-Uranus Encounter: 1. SPAN Accelerated Data Transfer to Remote Voyager Scientists, 2. PDS Increased Quick-Look Analysis for Voyager Scientists," *Information Systems Newsletter* 5:1.

Wulf, William A. 1989. "The National Collaboratory—A White Paper," Appendix A in *Towards a National Collaboratory*, the unpublished report of an invitational workshop held at the Rockefeller University, March 17-18, 1989 (Joshua Lederberg and Keith Uncapher, co-chairs).

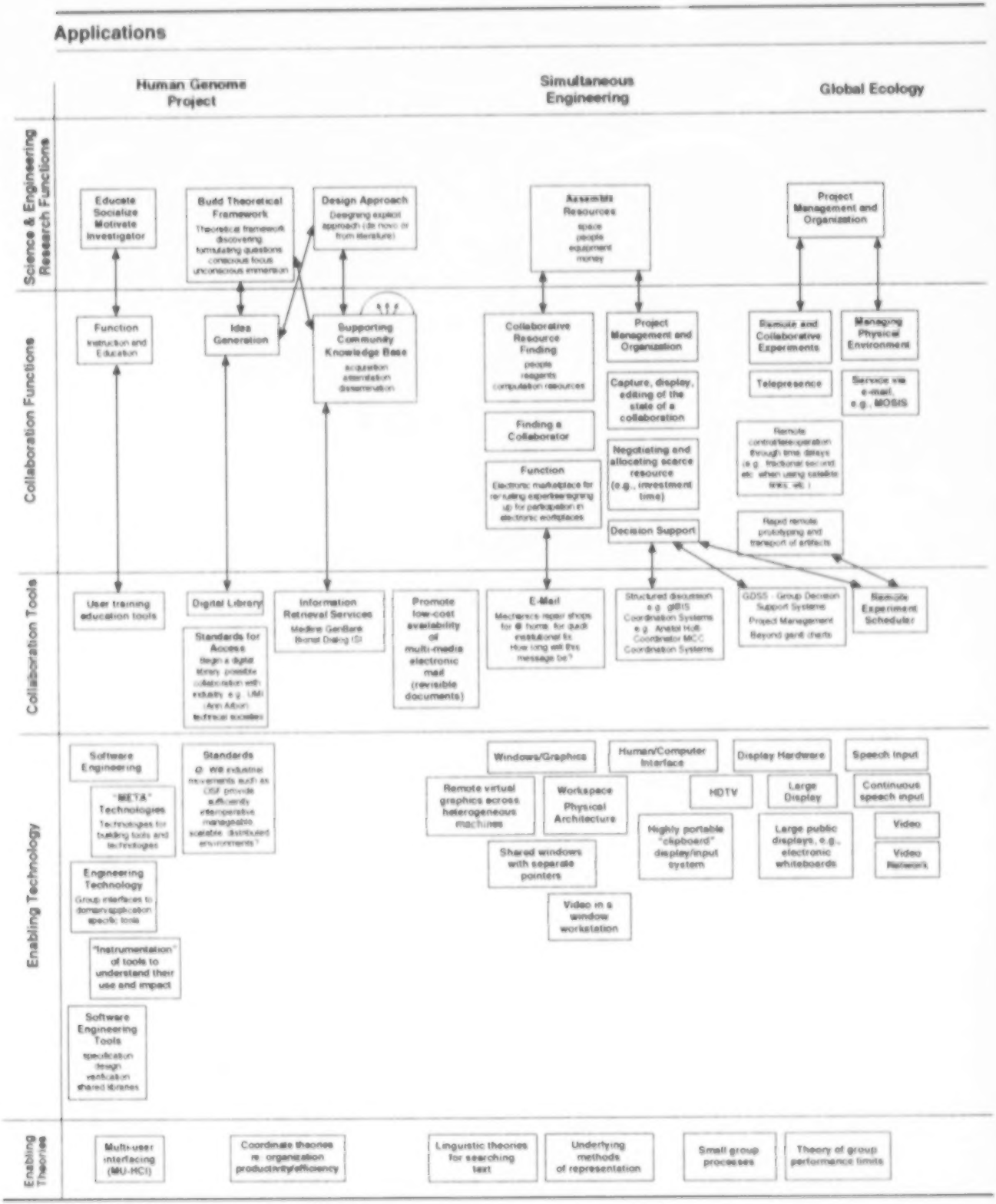
Appendixes

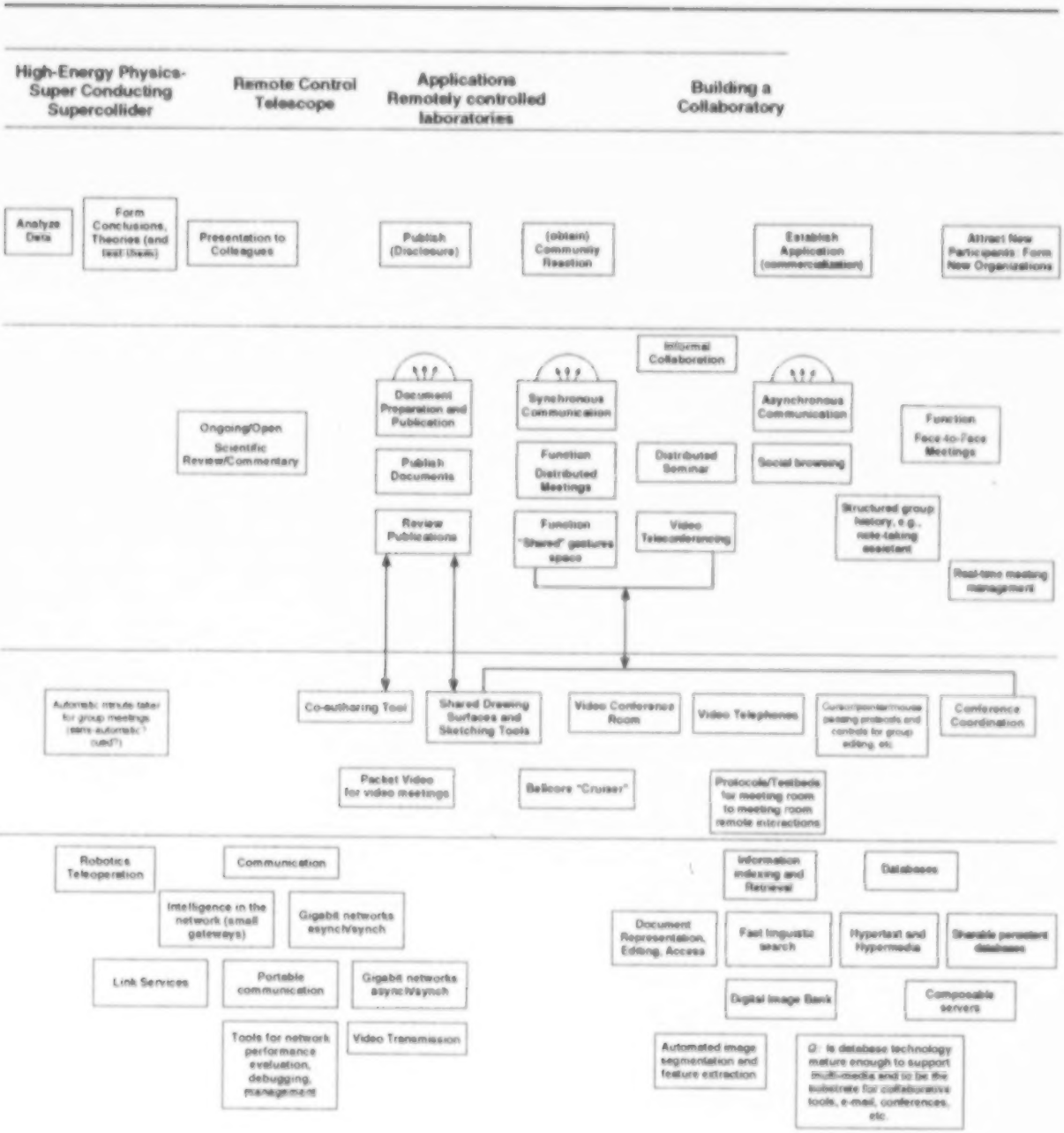
Appendix A

Elements of a Functional Collaboratory

The diagram reproduced overleaf shows the elements of a functional collaboratory—the kinds of connections between scientific and engineering applications and functions and the enabling technology and research objectives envisioned for a national collaboratory as discussed at an NSF-sponsored invitational workshop. The diagram organizes the discussion into levels. The top level is a sampling of applications. The next level is a breakdown of general scientific functions common to most scientific projects. An example of a scientific function is obtaining peer commentary. Different projects emphasize these functions to different degrees. The next level presents some collaborative functions. An example of a collaborative function is co-authoring a document. In general, these functions are typical of a number of different scientific disciplines. The next level describes some collaboration tools and systems that support one or more collaborative functions. The bottom two levels correspond to theories and technologies (infrastructure) that are used in the creation of tools. A program for a national collaboratory would include activities at all of these levels.

Prepared by Mark Stefik, Xerox Palo Alto Research Center, this diagram is reprinted from *Towards a National Collaboratory*, the unpublished report of an invitational workshop held at Rockefeller University on March 17-18, 1989 (Joshua Lederberg and Keith Uncapher, co-chairs).


Applications
Human Genome Project
Simultaneous Engineering
Global Ecology
Science & Engineering Research Functions
Educate Socialize Motivate Investigator
Build Theoretical Framework
Theoretical framework discovering formulating questions conscious focus unconscious immersion
Design Approach
Designing explicit approach (de novo or from literature)
Assemble Resources
space people equipment money
Project Management and Organization
Collaboration Functions
Function
Instruction and Education
Idea Generation
Supporting Community Knowledge Base
acquisition orientation dissemination
Collaborative Resource Finding
people reagents computation resources
Project Management and Organization
Capture, display, editing of the state of a collaboration
Finding a Collaborator
Negotiating and allocating scarce resource (e.g., investment time)
Function
Electronic marketplace for recruiting experts/engaging up for participation in electronic workplaces
Decision Support
Remote and Collaborative Experiments
Managing Physical Environment
Telepresence
Service via e-mail, e.g., MOSIS
Remote control/teleoperation through time delays (e.g. Mars/lunar second, etc. when using satellite links, etc.)
Rapid remote prototyping and transport of artifacts
Collaboration Tools
User training education tools
Digital Library
Standards for Access
Information Retrieval Services
Medline GenBank Biomat Dialog ID
Promote low-cost availability of multi-media electronic mail (revisible documents)
E-Mail
Mechanics repair shops for @ home, for audi instational fix How long will this message be?
Structured discussion e.g. gIBIS Coordination Systems e.g. Amatol Hub Coordinator MCC Coordination Systems
GDSS - Group Decision Support Systems Project Management Beyond gant charts
Remote Experiment Scheduler
Enabling Technology
Software Engineering
"META" Technologies
Technologies for building tools and technologies
Standards
Engineering Technology
Group interfaces to domain/application specific tools
"Instrumentation" of tools to understand their use and impact
Software Engineering Tools
specification design verification shared libraries
Windows/Graphics
Human/Computer Interface
Display Hardware
Speech Input
Remote virtual graphics across heterogeneous machines
Workspace Physical Architecture
HDTV
Large Display
Continuous speech input
Highly portable "clipboard" display/input system
Large public displays, e.g., electronic whiteboards
Video
Video Teleworks
Shared windows with separate pointers
Video in a window workstation
Enabling Theories
Multi-user interfacing (MU-HCI)
Coordinate theories re organization productivity/efficiency
Linguistic theories for searching text
Underlying methods of representation
Small group processes
Theory of group performance limits


High-Energy Physics-
Super Conducting
Supercollider
Remote Control
Telescope
Applications
Remotely controlled
laboratories
Building a
Collaboratory
Analyze Data
Form Conclusions, Theories (and test them)
Presentation to Colleagues
Publish (Disclosure)
(obtain) Community Reaction
Establish Application (commercialization)
Attract New Participants Form New Organizations
Informal Collaboration
Document Preparation and Publication
Synchronous Communication
Asynchronous Communication
Function Face-to-Face Meetings
Ongoing/Open Scientific Review/Commentary
Publish Documents
Function Distributed Meetings
Distributed Seminar
Social browsing
Structured group history, e.g., note-taking assistant
Review Publications
Function "Shared" gestures space
Video Teleconferencing
Real-time meeting management
Automatic minute taker for group meetings (semi-automatic? useful?)
Co-authoring Tool
Shared Drawing Surfaces and Sketching Tools
Video Conference Room
Video Telephones
Conference Coordination
Packet Video for video meetings
Bellcore "Cruiser"
Protocols/Testbeds for meeting room to meeting room remote interactions
Robotics Teleoperation
Communication
Information Indexing and Retrieval
Databases
Intelligence in the network (small gateways)
Gigabit networks esynch/esynch
Document Representation, Editing, Access
Fast linguistic search
Hypertext and Hypermedia
Shareable persistent databases
Link Services
Portable communication
Gigabit networks asynch/asynch
Digital Image Bank
Composable servers
Tools for network performance evaluation, debugging, management
Video Transmission
Automated image segmentation and feature extraction
Q: Is database technology mature enough to support multi-media and to be the substrate for collaborative tools, e-mail, conferences, etc.

Appendix B

Workshop Programs and Participants

MOLECULAR BIOLOGY WORKSHOP

Program

Thursday, March 12, 1992

8:30 a.m. Continental Breakfast

9:00 a.m. Welcome and Introduction—Alastair Cameron and Bruce Schatz
- Background of the collaboratory concept—Need to leverage limited human, financial, and time resources and enable collaboration among scientists
- Workshop goals
  - New ideas for collaboration and the technology that will facilitate collaboration
  - Priorities and agenda for development of collaboratories
- Introductions and personal statements about perspectives on collaboration

9:30 a.m. Current Issues and Trends in Molecular Biology—David Kingsbury
- Communities and problem domains
- Types of available knowledge: genes, maps, sequences, literature
- Types of available databases: formal, informal, collaborative, laboratory

10:00 a.m. Molecular Biology and Computer Technology: Lessons Learned from Previous Projects
- GenBank/Bionet—Douglas Brutlag
- Genome Data Base—Robert Robbins

10:45 a.m. Break

11:00 a.m. Current Molecular Biology Collaboratories
- Informal Community Systems—Bruce Schatz
- Archival Library Systems—Jim Ostell

12:00 noon Working Lunch with Demonstrations of the Worm Community System and of the National Center for Biotechnology Information Portable Core Library

1:30 p.m. Current Issues and Trends in Computer Technology: Information Infrastructure
- Hardware and Networks—Vint Cerf

- Software and Databases—Nat Goodman

2:00 p.m. Impediments to Adoption of Collaboration Technology—Lee Sproull
- Perceived value of technology to the molecular biology community
- User readiness and acceptance of the technology
- Training for computational biologists
- Lack of institutional and information infrastructure
- Lack of critical mass of on-line materials
- Lack of financial and institutional support for large projects

3:00 p.m. Brainstorming
- How might collaboration aid molecular biology and molecular biologists?
- What existing problems in molecular biology might collaboratories solve or ameliorate?
- Given sufficient funding, what would you want to build?
- Which communities are the most likely to be initial candidates for collaboratories?
- How will these collaboratories benefit the scientists who use them?

5:00 p.m. Reception

6:00 p.m. Dinner

Friday, March 13, 1992

9:00 a.m. Welcome and Statement of Purpose for Day Two—Bill Wulf
Recap of Brainstorming Session—Alastair Cameron and Bruce Schatz

9:30 a.m. How do we get there from here?
- What must be done to realize collaboratories in molecular biology in the near term?
- What specific proposals might be made?
- What tools do molecular biologists use that could be shared with scientists of other disciplines?
- How can the needs of molecular biologists be leveraged with the needs of other scientists to establish collaboratories? What other scientific communities might be targeted?

10:45 a.m. Break

11:00 a.m. What are the unsolved problems?
- What will they keep us from achieving?
- What must be done in the long term to address these problems?

12:30 p.m. Working Lunch

1:30 p.m. Conclusions and Report Planning

2:30 p.m. Summary and Adjourn

Participants

Douglas Brutlag, Stanford University School of Medicine
Alastair Cameron, Harvard College
David J. Galas, Department of Energy
Nat Goodman, Whitehead Institute
Steven Hilgartner, Columbia University
Tim Hunkapiller, California Institute of Technology
David Kingsbury, George Washington University Medical Center
Jim Ostell, National Library of Medicine
Ross Overbeek, Argonne National Laboratory
Robert Robbins, Johns Hopkins University
Laurence Rosenberg, National Science Foundation
Bruce Schatz, University of Arizona
Cassandra Smith, University of California, Berkeley
Gio Wiederhold, Defense Advanced Research Projects Agency
John Wooley, National Science Foundation

OCEANOGRAPHY WORKSHOP

Program

Thursday, April 23, 1992

8:00 a.m. Continental Breakfast

8:30 a.m. Welcome and Introduction—Vint Cerf and Tom Dickey
- Background of the collaboratory concept—Need to leverage limited human, financial, and time resources and enable collaboration among scientists
- Workshop goals
  - New ideas for collaboration and the technology that will facilitate collaboration
  - Priorities and agenda for development of collaboratories
- Introductions and personal statements about perspectives on collaboration

9:00 a.m. Current Working Collaboratories
- Modeling Systems in Oceanography—Lew Rothstein
- Informal Community Systems in Molecular Biology—Bruce Schatz

10:00 a.m. Break

10:15 a.m. Oceanography and Computing: Needs and Desires of Oceanographic Modelers Discussion—Vint Cerf
- What are the needs of the modeling community with respect to the technology?
- What technology is desirable and why?
- How will the community be enabled through increased connectivity and/or additional computer resources?
- What are some possible pilot projects?

12:15 p.m. Working Lunch with Demonstrations of Oceanographic Modeling Systems and Informal Community Systems

1:15 p.m. Obtaining Access to Field Data: Lessons From Current Practice (10-minute presentations with discussion)
- Identifying the Needs—Tom Dickey
- Near-Real-Time Data Acquisition: The Atlas Moorings Pilot—Ants Leetmaa and Mike McPhaden
- High-Frequency Telemetry Solutions—Mel Briscoe
- Sierracom—Phil Walker
- Ship-to-Land Communication—Andy Maffei

2:15 p.m. Break

3:00 p.m. Day-to-Day Collaboration Among Oceanographers: Problems and Solutions (10-minute presentations with discussion)
- Multidisciplinary Collaboration—Peter Wiebe
- Omnet and SCIENCEnet—Bob Heinmiller

4:00 p.m. Impediments to Collaboration—Discussion

- Access to data
- Cultural limitations
- Geographic and physical limitations
- Perceived value of technology to the oceanography community
- User readiness and acceptance of the technology
- Training/support for computational oceanography
- Lack of institutional and information infrastructure such as laboratories and equipment

5:00 p.m. Reception

6:00 p.m. Dinner

Friday, April 24, 1992

8:00 a.m. Continental Breakfast

8:30 a.m. Statement of Purpose for Day Two and Recap of Thursday's Discussions—Vint Cerf

9:00 a.m. Brainstorming

- How might collaboration aid oceanography?
- What existing problems in oceanography might collaboratories solve or ameliorate?
- Given sufficient funding, what would you want to build?
- Which communities are the most likely to be initial candidates for collaboratories?
- How will these collaboratories benefit the scientists who use them?

10:30 a.m. Break

10:45 a.m. How do we get there from here?

- What must be done to realize collaboratories in oceanography in the near term?
- What specific proposals might be made?
- What tools do oceanographers use that could be shared with scientists of other disciplines?
- How can the needs of oceanographers be leveraged with the needs of other scientists to establish collaboratories? What other scientific communities might be targeted?

12:30 p.m. Working Lunch

1:30 p.m. What are the unsolved problems?

- What will they keep us from achieving?
- What must be done in the long term to address these problems?

2:30 p.m. Conclusions and Report Planning

3:30 p.m. Adjourn

Participants

Jim Baker, Joint Oceanographic Institutions Inc.
Melbourne Briscoe, Office of Naval Research
Vint Cerf, Corporation for National Research Initiatives
Alan Davis, Florida State University
Tom Dickey, University of Southern California
David Evans, Office of Naval Research
Robert Heinmiller, Omnet Inc.
Ellen S. Kappel, Joint Oceanographic Institutions Inc.
Gary Koob, Office of Naval Research
Richard Lambert, National Science Foundation
Ants Leetmaa, NOAA Climate Analysis Center
Andrew Maffei, Woods Hole Oceanographic Institution
Mike McPhaden, NOAA Pacific Marine Environmental Laboratory
Rebecca G. Moser, Joint Oceanographic Institutions Inc.
Lew Rothstein, University of Rhode Island
Bruce Schatz, University of Arizona
Philip Walker, Sierracom
Peter Wiebe, Woods Hole Oceanographic Institution
Stan Wilson, NOAA National Ocean Service

SPACE PHYSICS WORKSHOP

Program

Thursday, July 9, 1992

8:00 a.m. Breakfast

8:30 a.m. Welcome and Introduction—Vint Cerf and C.T. Russell
- Background of the Collaboratory Concept—Need to leverage limited human, financial, and time resources and enable collaboration among scientists
- Workshop Goals
  - New ideas for collaboration and the technology that will facilitate collaboration
  - Priorities and agenda for development of collaboratories
- Introductions and Personal Statements About Perspectives on Collaboration

9:00 a.m. Early Collaborations
- Space Physics Analysis Network: The Early Years—Jim Green
- Coordinated Data Analysis Workshops—Dan Baker
- Atmospheric Explorer/Dynamics Explorer Data System—Dave Winningham

10:00 a.m. Break

10:15 a.m. Recent Collaborative Efforts
- Cometary Studies on Giotto—Marcia Neugebauer
- Magnetospheric Studies on AMPTE—Steve Fuselier
- Remote Operation of Instrumentation—John Kelly

12:00 noon Lunch

1:00 p.m. Future Collaborative Efforts
- Geospace Environment Modeling Program—Tim Eastman
- Space Physics Data System—Dave Winningham
- International Solar-Terrestrial Program Mission—Dan Baker

2:00 p.m. Access to Data
- Policy on the Access to Publicly Funded Data—Jim Willett
- State of the NSSDC Archives Master Directory—Jim Green
- Experiences of a Scientist—Bob McPherron

3:00 p.m. Break

3:15 p.m. Electronic Networks in Collaborative Studies
- Numerical simulations
- Theoretical studies
- Experimental studies

4:15 p.m. Discussions of Impediments to Collaboration—All

5:00 p.m. Reception

6:00 p.m. Dinner

Friday, July 10, 1992

8:00 a.m. Breakfast

8:30 a.m. Statement of Purpose for Day Two and Recap of Day One—Vint Cerf

9:00 a.m. Brainstorming on Contents of Report
- How might collaboration aid space physics?
- What existing problems in space physics might collaboratories solve or ameliorate?
- Given sufficient funding, what would you want to build?
- Which communities are the most likely to be initial candidates for collaboratories?
- How will these collaboratories benefit the scientists who use them?

10:30 a.m. Break

10:45 a.m. How do we establish the needed infrastructure?
- What must be done to realize collaboratories in space physics in the near term?
- What specific proposals might be made?
- What tools do space physicists use that could be shared with scientists of other disciplines?
- How can the needs of space physicists be leveraged with the needs of other scientists to establish collaboratories? What other scientific communities might be targeted?

12:00 noon Lunch

1:00 p.m. What are the unsolved problems?
- What will they keep us from achieving?
- What must be done in the long term to address these problems?

2:30 p.m. Conclusions and Report Planning

3:30 p.m. Adjourn

3:30-
5:00 p.m. Executive Committee Meeting

Participants

Daniel Atkins, University of Michigan
Daniel Baker, NASA Goddard Space Flight Center, National Space Science Data Center
Joseph Bredekamp, NASA Headquarters
Vint Cerf, Corporation for National Research Initiatives

Timothy Eastman, National Science Foundation
Steve Fuselier, Lockheed Palo Alto Research Laboratory
James Green, NASA Goddard Space Flight Center, National Space Science Data Center
John D. Kelly, Stanford Research Institute International
William Kurth, University of Iowa
Barry M. Leiner, Universities Space Research Association
Janet G. Luhmann, University of California, Los Angeles
Robert L. McPherron, University of California, Los Angeles
Marcia Neugebauer, Jet Propulsion Laboratory
C.T. Russell, University of California, Los Angeles
James Willett, NASA Headquarters
John D. Winningham, Southwest Research Institute

Appendix C

Rules Governing Access to and Use of the CDAW-9 Database

Certain projects operate under a set of rules (e.g., the Rules of the Road of the Dynamics Explorer and the International Sun-Earth Explorer Missions) which identify obligations of investigators and researchers with regard to data provided by other investigators. Because the activities of and principal publication by members of the Coordinated Data Analysis Workshop (CDAW) are, by design, collaborative efforts which greatly extend this interchange of data within the research community, the following rules were adopted for CDAW-9. If a conflict were identified between these rules and the rules governing a data set provided to CDAW-9, the flight project rule would be adopted and an amended copy of the CDAW-9 rules forwarded to each member.

1. Access to the CDAW-9 database will be granted to established, participating members of CDAW-9. Members who withdraw from participation in CDAW-9 are obligated to continue to respect the rules established here.

   CDAW members are those scientists belonging to institution-specific research teams providing data to the CDAW-9 database, and/or belonging to modeling/theory teams whose participation is invited by the CDAW Program Committee. CDAW members are also those individuals who, through regular attendance at regularly scheduled workshops, make significant contributions to organizing, reformatting, and interpreting data in the database.

2. Members of CDAW-9 may share data with members of their research team, but are not entitled to further distribute data provided by other investigators. The CDAW-9 member is responsible to CDAW-9 for ensuring that these rules are followed by members of his/her research team.

3. The CDAW-9 database is established solely for correlative studies by members of the CDAW. Access to the database and support software provides individual members of CDAW the means of correlating the established data sets. Preparation of an established data set is the responsibility of the investigator who supplies the data, and is not to be undertaken by other members of the CDAW.

4. When an investigator's data are used in the analysis of an event, the investigator responsible for providing these data should be kept informed of what they are being used for, should be invited at an early and appropriate time to participate in the correlative analysis, must be invited to participate as a co-author of a research paper or abstract, and must be given a reasonable oppor-

---

NOTE: Reprinted from unpublished material developed internally for use by participants in Coordinated Data Analysis Workshop-9, one of a series of voluntary programs sponsored by the National Aeronautics and Space Administration.

tunity to review that work prior to submission for publication. If an investigator chooses to decline co-authorship after reviewing the work, that investigator is still encouraged to provide a critical review to the principal author if a disagreement exists concerning the interpretation of that investigator's data.

5. If an investigator declines co-authorship of a paper, an acknowledgement of that investigator's participation shall be provided. Authors are reminded that acknowledgements of the joint efforts of CDAW participants, the National Space Science Data Center (NSSDC), and of the Space Physics Analysis Network (if appropriate) assist in making these efforts more visible to the research community and supporting agencies, and assist in establishing and enhancing future joint activities.

6. New CDAW members will be welcome at any time, and in particular after the first major (series of) publication(s) of CDAW-9 results. These new members will abide by these rules of the road. Unrestricted access to the database will be granted no later than May 1994, and possibly earlier if the Program Committee, after polling the members, determines to do so. Investigators will be given the opportunity to withdraw their data from the CDAW-9 database prior to this public release.

7. Submission of data to the NSSDC for CDAW-9 by an investigator does not constitute submission of data to the NSSDC as part of any contractual obligation.

8. Summary plots of DE-1 and -2 data and of DE-1 and Viking auroral imaging survey slides are provided to identify scientifically interesting periods. They are designed to support data analysis efforts only, and are not to be used for any other purpose without the specific authorization of the appropriate Principal Investigators. The DE-1 and -2 summary plots contain unverified data.

9. The DE-1 and Viking auroral imagery Principal Investigators have reserved the exclusive right to pursue magnetic conjugacy studies with their imaging data, for the time being. Thus, such studies are "off limits" to CDAW-9 members until the relevant Principal Investigators waive this right. It is expected that such a waiver will occur in the not-too-distant future.

Appendix D

Training Computational and Mathematical Biologists

INTRODUCTION

It has been estimated that in mid-1990, there were approximately 4000 professional-level scientists identifiable as computational or mathematical biologists. These scientists were found in a wide variety of institutions and in a wide range of positions within those institutions.

The pattern of distribution of these individuals among and within different institutions appears to be related to their academic training. For example, mathematicians and computer scientists who have primarily followed an interest in the biological sciences generally work as biologists and find themselves in nonacademic research positions in industry, government or private research institutes, or quasi-academic research centers (e.g., supercomputer centers). A small minority are in biology departments. In contrast, mathematicians who have continued to pursue research activities in mathematics, choosing biologically related problems or examples, or collaborating with biologists, tend to remain in departments of mathematics or applied mathematics in academic institutions. Computer scientists follow a similar pattern. Statisticians may be found in statistics departments, biostatistics groups or departments, or even in biological sciences departments, depending on the extent of their involvement with biological problems, and the local structure of the institutions within which they work.

Biologists who rely on computational and mathematical tools in their research activities are found in many institutions. A large number have moved into industry where they play a role in the analysis of macromolecules in biotechnology and pharmaceutical companies. Another major source of employment is in government and private research institutes, which tend to focus on problem-oriented research and directly utilize their computational biology skills. In the academic environment, computational biologists pursuing accepted biological problems are found in a variety of departments of biology (including departments with related names such as genetics, ecology, and evolutionary biology, molecular biology, and microbiology), chemistry, and biochemistry.

The character of the institutional acceptance of these interdisciplinary activities depends on two factors: the need of the institution for problem-oriented work, and the traditional academic expectations for the performance of the individual. For example, biology departments place their emphasis on disciplinary achievements, and computational and mathematical approaches are secondary to the disciplinary results. Therefore, the infusion of mathematical and computational tools is dependent on the confidence of researchers that they can afford to invest the time and effort to enable them to use this approach, let alone develop new tools. Thus in many cases, computational and mathematical biology makes a back-door entrance into the academic world. In contrast, these approaches are embraced more

---

NOTE: Reprinted from Appendix 3 of the final report of an NSF-sponsored workshop, Training Computational and Mathematical Biologists, held at the Banbury Center of the Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, December 9-11, 1990.

directly by industry and research institutes whose problem-oriented programs utilize a broader range of approaches, including direct application of mathematical and computational techniques.

The workshop participants' assessment is that in the immediate future, this situation will not undergo a substantial change. Therefore, scientists expecting to enter the academic research world will continue to need a strong disciplinary grounding for their cross-disciplinary work. Employment opportunities in industry and research institutes appear to be stable, or growing slowly. Such centers will continue to be major sites for the development of computational techniques and applications in biology.

Because of their frequently strong mathematical and computational environments, and the less frequent presence of rigid departmental structures, one possible source of future growth for computational biology is the four-year college. Mathematical and computational approaches fit well within the research environments found in these institutions, and they are likely to find effective implementation in the teaching programs. In this context, faculty in these institutions may be expected to employ mathematical and computational techniques in both research and the development of teaching aids that will eventually find their way into research institutions. However, here again, strong disciplinary training will be essential as the basis for the research approach.

PROFILES OF COMPUTATIONAL AND MATHEMATICAL BIOLOGISTS

In the past, most of the migration of scientists into computational biology has been from disciplines outside of biology (e.g., math, physics, chemistry, computer science, etc.). Physicists become biologists, but not the reverse. This migration and its asymmetry have been prompted by successful application of domain-specific technology to solving biological problems.

Many early successes in computational biology were obtained by scientists who were primarily biologists with marginal skills in computer science and mathematics (programming skills and some algorithmics), while many others were the result of work by scientists with extensive mathematical and computational backgrounds. However, as the problems under investigation become more complex, training which provides great depth in quantitative analysis will be essential.

Current interest and excitement in computational and mathematical biology are driven in large part by neurobiology, global change, and genomics. In all of these areas, vast amounts of information are accumulating at a rate that precludes human absorption and, hence, understanding. Biology needs tools for manipulating and analyzing information. In order for training environments to be maximally effective, there must be a clear understanding of which professional profiles are suitable for current and future researchers in computational and mathematical biology.

The profiles which follow are dependent upon the nature of the position. Academicians tend to reside within traditional departmental units, whereas in industrial settings and research institutes, there is a wider range in the mixtures of disciplines in working groups. The following lists of specialties within computer science, mathematics, and biology are those in which there is substantial research activity today and where there is likely to remain some research focus in the future.

Computer Scientists

Most computer scientists retain their primary professional identification with computer science. They tend to view biological applications as a source of computer science problems. Biological applications are new to computer scientists, and the traditions across the interface are developing at a moderate pace. The tendency is to cross the line as a senior scientist by developing collaborations. There are some successful scientists in this field whose first exposure to biology was at the graduate level. Examples of the areas of computer science in which such collaborations take place are:

- Artificial neural networks (AI)
- Algorithmics
- Database design and theory
- Visualization (graphics)

Biologists

Biologists working on computational problems come from a plethora of backgrounds: computer science, mathematics, statistics, engineering, physics, and chemistry as well as biological disciplines. The biological sciences are themselves diverse, and different areas of biology draw upon very different quantitative skills. Those biologists who have crossed the boundaries between biology and other disciplines have often done so to address specific biological problems. Their acceptance by the biological community has been out of necessity, since many biological problems require technology that has been driven by insight and intuition from other disciplines. This report [on training computational and mathematical biologists] is motivated by the assumption that this trend will accelerate in the near future in areas such as genomics, neurobiology, imaging, structural biology, and issues of global climate change. Many of these developments have been initiated by scientists whose initial training was outside biology (e.g., mathematics, chemistry, and physics). The current technological advances will require a new range of quantitative skills beyond the norm of current curricula in the biological sciences. Biological sciences that currently draw substantially from the computational and mathematical sciences include:

- Population biology, including ecology and genetics
- Molecular biology
- Molecular genetics
- Cellular biology
- Neurobiology
- Biophysics and structural biology
- Ecosystem ecology
- Epidemiology
- Physiology

Mathematicians

There is a long tradition of mathematicians and statisticians working on biological problems. Indeed, the field of statistics grew largely out of biological origins, and there is a substantial portion of the statistics community working on problems of biometry and biostatistics. There is also a small but stable community of mathematical biologists working within departments of pure and applied mathematics. Some members of this community migrate to biological departments during the course of their careers, while others remain in mathematical science departments. Those who do remain within mathematical science departments either establish a career based upon collaborations with biologists, or focus upon mathematical questions driven by biological problems. In some cases, threads of mathematical research initiated by biological problems take on a life of their own as interesting areas of mathematics per se. Areas of mathematics making substantial contributions to biology include:

- Applied mathematics (differential equation models, image processing and analysis)
- Probability (sequence analysis, interacting particle systems)
- Statistics

- Discrete mathematics
- Topology and differential geometry

SUMMARY OF THE CURRENT STATUS

With regard to the current panorama of activity, we perceive that several difficulties exist. First, computer scientists are not sufficiently involved in computational biology. Their work is frequently on problems so abstracted from the application as to make them less than fully effective as collaborators. Another limitation is that biologists tend to view the work of computational scientists as service, and not original research, which tends to alienate this community. Mathematicians are caught between mathematical peers who evaluate their work on the basis of its mathematical depth and elegance, and biologists who have little appreciation for theory that does not have a direct bearing on the interpretation of experimental data. Finally, those biologists who have invested in cross-training are frequently misunderstood and undervalued by their colleagues, most of whom do not understand how to evaluate their work.

Computer science is a new discipline that is rapidly maturing. As the field develops, a tradition of interdisciplinary work will evolve much as it has for mathematics, especially statistics. This will, in part, alleviate the problem of computer scientists' involvement. A greater emphasis on the early grounding in scientific disciplines while at the undergraduate level should also help to cultivate computer scientists with a stronger interdisciplinary focus. As the needs for computation in the various areas described above become clearer, the biological community must become increasingly more tolerant and accepting of computational biologists within their midst. As a result of this and other factors, such as heavy dependence on physical measurement, the training of biologists at all levels must become increasingly more quantitative in nature.

ENCOURAGING INTERACTIONS

The most effective way to encourage interactions between mathematicians and computer scientists on the one hand, and biologists on the other, is through direct co-involvement with a particular problem. This applies at all levels from undergraduate through senior scientist. The ways in which this interaction may be encouraged depend on the level and direction of movement (math/computer science to biology or biology to math/computer science). At present, the pattern is generally unidirectional, with movement from mathematics or computer science into biology as the dominant paradigm. Significant changes in this state of affairs are likely to require substantial curricular changes based upon effective means of overcoming the apprehension of most biology students towards mathematics.

Interaction can be improved through a strengthening of mechanisms that already exist. However, one area deserves much greater emphasis than is now the case, and that is support of small research groups with a genuine interdisciplinary focus: within this, substantial support is needed for postdoctoral scientists. Support of small group research will develop critical mass in important areas, will help to foster and sustain collaborative research, and will provide a crucial home for individuals who are in the early stages of (what is now) a cross-disciplinary research career.

The most effective mechanisms for stimulating these fields vary by the level of a scientist's career stage as outlined below.

Senior Researchers (Tenured and Above)

Math/Computer science to biology: support for sabbaticals and, later, research in biology.

Biology to math/computer science: support for visits to math research groups to learn/update new technical areas.

Pretenure

Most mathematics and statistics Ph.D. students will start in untenured positions. Changing fields (or, at least becoming more interdisciplinary) at such an early stage is a very risky career move, particularly by individuals approaching a tenure decision. One way to ameliorate this situation is through a new focus on PYI-level type support (National Science Foundation Presidential Young Investigator) for promising people (prestigious competitive awards).

Postdoctoral

Support for postdoctoral training within existing grants is essential. Postdocs are an important educational component of existing research groups, and are very scientifically profitable in the short term. These grants should support a given individual for multiple years, and not be specifically tied to a particular investigator within the group. This mechanism allows quick response to changing areas of interest, while providing enough time for a postdoctoral fellow to develop a useful independent research focus.

Another aid to young investigators is the computational research associates program at the NSF-sponsored supercomputing centers. This program is of great value to the biological sciences, and the field would benefit from its continued existence. However, to be maximally effective these investigators must be part of an active and focused research program and not "generalists" in applied computer science.

The concepts behind these training programs are not based on the assumption that all people passing through them will eventually obtain tenure-track positions in universities.

Graduate Students

An important source of mathematical biologists comes from mathematically trained undergraduates who change fields early in their postgraduate education. Such students are then main-stream biologists with the requisite quantitative background to enter the fields of mathematical or computational biology. The educational challenge for students with this background is the continuation of the quantitative approach to biology in a supportive environment. This requires an appropriate mentor and an appropriate departmental or graduate group environment so that the student's background is valued and prior training reinforced. Given the many opportunities available to an undergraduate with computer science or mathematical training, it is essential that graduate student support be provided to entice these students to forego the immediate gratification of lucrative employment for the longer-term prospects of graduate training and research careers in biology. To this end the continued and renewed support of training grants or traineeships (for example, in the research groups described above) are of central and continuing importance.

Furthermore, educational institutions must be encouraged to recognize the need for training students in these areas as a means of dealing with the future of biological research. To this end institutional and departmental support of fellowships and RA (research assistant) positions is of supreme significance. Cross-training students at the graduate level will lengthen an educational process that already can be inordinately long. Freeing a student from the demands of a teaching assistantship or a research assistantship with responsibilities to further the work of a principal investigator will help make such programs educationally feasible. It would be especially appealing to find a mechanism to support mathematical or computational biologists within the structure of departments of mathematics or computer science.

One of the most significant factors in the training of graduate students is the role model of the major professor. This mentorship plays a greater role in the ultimate aspirations of a student than is generally acknowledged. The successes, failures, and frustrations of a student's mentor play a profound role in the expectations and aspirations of a student. In this context the small-group research environment is a highly significant environment in which to train students for the future of the biological sciences.

Undergraduate

In most institutions it is very common for the top biology students, especially those interested in eventual graduate study, to participate in undergraduate research projects, especially in their junior and senior years. This opportunity should not be confined to biology students, but should be expanded wherever possible to include interested students from mathematics and computer sciences whenever possible. The proper environment is essential to the nurturing of a student that might wish to commit to a career in the biological sciences, using this valuable undergraduate training. To this end the National Science Foundation REU (Research Experiences for Undergraduates) program provides an extraordinary opportunity in the math/biology area.

One area of extreme importance for the future development of a cadre of computational and mathematical biologists, and for the continued recruitment of students into biophysics and related disciplines, is the development of better course materials devoted to the quantitative approach to biology. The workshop participants valued very highly the concept of "enculturation of quantitative thought" through the introduction of quantitative approaches in biology courses.

Precollege

While there was considerable discussion during the workshop regarding the state of precollege science education, no specific recommendations were developed. Many private and government agencies have focused great attention on this problem, and it remains a top national priority. There was general agreement that two issues posed particular concern to the participants. First, the need to involve parents more fully in the educational process. This is particularly important in groups which do not have a cultural history of educational achievement. The second concern was the current selection of the "ultimate underachiever" as the folk hero of the nation's children. We believe that this message is alarmingly inappropriate in the current context of rapid technological change and global competition. The participants hope that the leadership of the Education and Human Resources Directorate of the National Sciences Foundation will use its influence and insight to find a mechanism to reverse this trend.

Summary Principles

1. If time is limited for education, spend it in mathematics, not computer science.*
2. What we want is an attitude/consciousness change, so that people are aware of the input of the "other" type of science in their own area.
3. While collaboration will enhance the science of the current generation, we are seeking to change the way that biology is done by changing the way biologists are educated for the next forty years.

FUNDAMENTAL EDUCATIONAL PRINCIPLES

Undergraduate Education

General Course Content

The cross-disciplinary aspects of modern science must be emphasized in all undergraduate science and mathematics courses. The role of computer science and mathematics, as well as technologies from physics and chemistry, need to be presented in biology courses. In contrast, the research areas that have used various tools of computer science and mathematics in the experimental sciences should be identified throughout mathematics and computer science courses.

Mathematics and Computer Science Majors

All mathematics and computer science majors should have required experimental science courses. We recommend a minimum of two years that can be concentrated in one area or spread over the basic sciences. The purpose of this is to provide the student with an understanding of the vocabulary and concepts and an experience of the ways in which mathematics or computer science have contributed to other disciplines.

Biological Sciences

In order to produce biological scientists who will be qualified to do modern research, we strongly recommend that the science curricula require four years of mathematics and/or computer science. Representative courses might include programming, theory of algorithms, probability and statistics, linear algebra, calculus, discrete mathematics, and numerical analysis.

Consequences

Failure to implement these recommendations at a minimal level will foreclose the future for many undergraduates majoring in biological sciences. This originates in the types of problems that are coming into existence and that are consistently more and more dependent on quantitative skills for their solution. Secondarily, lack of training in these quantitative areas will limit the questions that can be asked by an investigator, and may come to threaten an individual's levels of funding. We must remember that we

*Note that this statement is taken directly from the publication. The Computer Science and Telecommunications Board does not necessarily endorse this statement of principle.

are addressing the education of persons who will be in the pool for the next forty years. If education changes are not implemented, much of biology will fail to thrive.

The broad education that we are proposing also permits people to change their minds and acquire additional course work in another field, even late in their studies, without having to start from the beginning.

Our recommendations should not be construed to support any concept that presupposes a gender-specific bias in the ability to perform. It may be that a type of math/computer science anxiety will become apparent if our recommendations are instigated. In order to counter this, we propose that support groups, personal tutorials, study circles and other tools of encouragement, and enhanced performance/esteem be supported so that they are readily available.

ADDITIONAL RECOMMENDATIONS

Part of the difficulty in implementing the course recommendations may be the prevalence of "premed" education as a major component of biology curricula. Although there will be a number of additional consequences, it would be well worth considering the restructuring of the undergraduate major so that "premeds" follow a separate track and their presence does not determine the future of an academic discipline.

It is incumbent upon those who practice cross-disciplinary science and mathematics/computer science to become both role models and mentors for others. It is particularly important for representatives of under-represented groups to make an effort to encourage others.

Several members of the group have suggested that a new type of biology course should be developed. It would cover the elements of modern biology, but highlighting the contributions of other disciplines. The hope is that someone will be inspired to write a founding text, one that will change the field.

Graduate Education

Continue to create opportunities for cross-disciplinary work. National Institutes of Health programs in molecular biophysics and the National Science Foundation research training groups are examples of attempts to encourage this type of interaction.

One-on-one mentor/student relationships are not sufficient to maintain cross-disciplinary development. Direct support for cross-disciplinary efforts would help to break down the interdepartmental barriers that frequently exist. Seminar groups or other frequent interactions should be encouraged.

New graduate students (and postdocs) might acquire an elementary grounding in a new field through summer institutes or some other "crash course." The courses would be taught by highly interactive, expert, senior-level researchers. For example, a course in basic molecular biological concepts could include molecular biology, biochemistry, and molecular biophysics. Emphasis would be on the vocabulary and point of view, that is, how the science is done and what its assumptions are. For a course on computation in genetics, this material might include basic computer science concepts, e.g., files, databases, algorithms and their use, graphics, and statistics. The benefits of such a course could also be made available to more senior investigators.

WOMEN AND OTHER UNDER-REPRESENTED GROUPS

In high school, women represent a reasonable proportion, approximately 30-40 percent, of those students who are interested in the physical sciences and mathematics. Partitioning begins in college and is nearly finished by graduate school. Some disciplines within the biological sciences do have equivalent or even over-balanced representation by women. Increasing the level of course work in mathematics and computer science may be threatening to some of these women. In order to prevent this, specific actions may well be necessary. Similarly, for some students from other under-represented groups, it may be necessary to have additional courses available at the undergraduate level to improve the level of computational competence of entering students.

NATIONAL ACADEMY PRESS
The National Academy Press was created by the National Academy of Sciences to publish the reports issued by the Academy and by the National Academy of Engineering, the Institute of Medicine, and the National Research Council, all operating under the charter granted to the National Academy of Sciences by the Congress of the United States.

BEST COPY AVAILABLE

ISBN 0-309-03818-6